WARWICK
THE UNIVERSITY OF WARWICK

A Thesis Submitted for the Degree of PhD at the University of Warwick

Permanent WRAP URL:
http://wrap.warwick.ac.uk/90153

Copyright and reuse:
This thesis is made available online and is protected by original copyright.
Please scroll down to view the document itself.
Please refer to the repository record for this item for information to help you to cite it.
Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

warwick.ac.uk/lib-publications

Eyewitness Identification Performance on Lineups for Distinctive Suspects

by

Melissa Fay Colloff

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Psychology

University of Warwick, Department of Psychology

December 2016

Table of Contents

List of Tables

List of Figures

Acknowledgements

I am extremely grateful to all those people who have supported and encouraged me throughout the course of my PhD. Primarily, I would like to thank Kimberley Wade, whose tireless guidance and support have been fundamental to both my research and future career. She is enthusiastic, honest, and always willing to go the extra mile to help her students. Kim is not only an outstanding supervisor, but is also an inspirational female scientist. I have learnt so much from her.

I also express my deepest gratitude to John Wixted for welcoming me into his lab and dedicating so much of his time to teach me signal detection statistics and modelling. My trip to San Diego was transformational and John's wisdom and willingness to continue to educate me are invaluable.

Many thanks also to Elizabeth Maylor, Neil Stewart and Laura Mickes: Elizabeth for her unrivalled proof-reading, astute feedback, and ageing expertise; Neil Stewart for giving his valuable time to help me to programme my studies; and Laura for insightful discussions about ROC analysis. I am so grateful to Heather Flowe for believing that I was able to undertake a PhD and providing me with many rewarding and worthwhile opportunities along the way.

I must also acknowledge the Psychology Department at Warwick University for supporting me financially and allowing me to be surrounded by a group of inspirational academics. Thank you also to all those people who made working in the office so enjoyable, and my lab mates—Sophie, Divya, and Harriet—for enduring countless practice presentations and draft papers. What a team!

Last, but certainly not least, I would like to thank my friends and family: my friends for humouring me and having faith that one day, I will be "Dr Colloff" (fingers crossed!), and my parents, sister, and Chris who have strived to give me every opportunity in life and have encouraged me throughout. I couldn't have done this without you. Thank you, Mom, for reading every word.

Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. The work presented (including data generated and data analysis) was carried out entirely by the author.

Inclusion of Published Work

Parts of this thesis have been published by the author.

Chapter 2 includes the following publication:

> Colloff, M. F., Wade, K. A., & Strange, D. (2016). Unfair lineups make witnesses more likely to confuse innocent and guilty suspects. *Psychological Science, 27,* 1227–1239. doi:10.1177/0956797616655789

Dr Kimberley Wade contributed to the planning of this research and Dr Kimberley Wade and Dr Deryn Strange provided feedback on drafts of the manuscript.

Chapter 3 includes the following publication:

> Colloff, M.F., Wade, K. A., Wixted, J.T., & Maylor, E. A. (in press). A signal-detection analysis of eyewitness identification across the adult lifespan. *Psychology and Aging.* doi:10.1037/pag0000168

Dr Kimberley Wade and Professor Elizabeth Maylor contributed to the planning of this research and Professor John Wixted assisted with the model fitting. All co-authors provided feedback on drafts of the manuscript.

In addition, Chapter 1, "Measuring discriminability", paragraphs 2–4 and Figure 1.4, appear in Colloff et al. (2016). Chapter 1, "Gauging the likely accuracy of identifications", paragraph 2, appear in Colloff et al. (in press).

Abstract

When constructing lineups for suspects with distinctive facial features (e.g., scars, tattoos, piercings), current police guidelines in several countries state that the distinctive suspect must not stand out. To this end, police officers sometimes artificially replicate a suspect's distinctive feature across the other lineup members (replication); other times, they conceal the feature on the suspect and conceal a similar area on the other members by pixelating the area (pixelation), or covering the area with a solid rectangle (block). Although these three techniques are used frequently, little research has examined their efficacy. This thesis investigates how the lineup techniques for distinctive suspects influence eyewitness identification performance and, in doing so, tests the predictions of a new model of eyewitness decision-making—the diagnostic-feature-detection model (Wixted & Mickes, 2014).

The research uses a standard eyewitness identification paradigm and signal detection statistics to examine how replication, pixelation, and block techniques influence identification performance: [1] compared to doing nothing to stop the distinctive suspect from standing out; [2] in young, middle-aged and older adults; and [3] when the culprit does not have the feature during the crime. It also examines [4] how variation in the way the suspect's feature is replicated influences identification performance.

The results converge to suggest that all three lineup techniques currently used by the police to accommodate distinctive suspects are equally effective and, when the culprit has the feature at the time of the crime, all enhance people's ability to discriminate between innocent and guilty suspects more than doing nothing to prevent a distinctive suspect from standing out. All three lineup techniques enable people of all ages to make highly confident decisions when they are likely to be accurate. These findings align with the predictions of the diagnostic-feature-detection model, which suggests that the model remains a viable theory of eyewitness decision-making.

Chapter 1 :
Introduction

Suppose that you were an eyewitness to a criminal event. Perhaps you saw a suspicious man in an area where you later learn a child has been abducted, or perhaps you caught a glimpse of the young man when he grabbed the bag from your arm. Because you have seen the culprit, you are a valuable source of information for the police officers investigating the case (Kebbell & Milne, 1998). The investigating officers ask you to describe the culprit, and then, sometime later, perhaps days, but perhaps months, they ask you to attempt to identify the culprit from a lineup. In this lineup, the police officers place their suspect (who is either innocent or guilty) among other known-to-be-innocent lineup members, called foils. If you make a positive identification of the police suspect—you say: "That's him!"—then this is likely to be interpreted as compelling evidence of guilt. Suspects who have been positively identified are more likely to be charged with the offence (Davis, Valentine, Memon, & Roberts, 2015; Flowe, Mehta, & Ebbesen, 2011) and are more likely to be found guilty at court (Devlin, 1976; Pozzulo, Lemieux, Wells, & McCuaig, 2006; Pozzulo, Lemieux, Wilson, Crescini, & Girardi, 2009) than those who have not. In short, eyewitness identification evidence plays a critical role in how a case proceeds through the Criminal Justice System.

The influence of eyewitness identification evidence in the Criminal Justice System is, however, concerning when we consider that memory can be unreliable. Eyewitness identification errors are frequent and can have profound consequences. In the United States since 1989, 344 convictions have been overturned on the basis of new DNA evidence. The Innocence Project estimates that over 70% of these wrongful convictions involved an eyewitness identification error—that is, the witness identified an innocent suspect (Innocence Project, 2016). Moreover, another type of identification error, failing to identify the culprit when he is in the lineup, can result in the real culprit being free to commit additional crimes. Subjects in experimental studies fail to identify a previously seen person about 50% of the time when that person is in the lineup (Wells, Memon, & Penrod, 2006), and when real witnesses in field studies choose from the lineup, they incorrectly identify a known innocent foil around 30% of the time (see Wells, Steblay, & Dysart, 2015 for a

discussion). Although it is impossible to know the number of occasions in which a real culprit has been falsely acquitted because a witness failed to positively identify him, given these statistics, it appears that this type of error is also likely to be common.

Experimental research

As a result of the frequency of eyewitness identification errors, psychological scientists have conducted experimental studies in the lab to examine the factors that may enhance or impair eyewitness identification accuracy (Wright, 2006). In these studies, researchers usually employ a mock crime methodology. Subjects watch a staged crime—sometimes live, but usually one that has been videotaped—then, after a delay, are presented with a lineup and have to attempt to identify the culprit. In real life criminal investigations the ground-truth is unknown, because police officers can never be certain if their suspect is innocent or guilty. But in lab studies, this factor can be controlled and experimentally manipulated. Some subjects are presented with a lineup in which the real culprit is present (a target-present lineup), and the remaining subjects are presented with a lineup in which the real culprit is absent (a target-absent lineup). Target-present lineups represent the real world situation in which the police suspect is guilty, whereas target-absent lineups represent the situation in which the police suspect is innocent. In both target-present and target-absent lineups, subjects can make one of three possible identification responses: they can identify the suspect, they can identify a foil, or they can reject the lineup and state that the real culprit is not present (see Table 1.1). In a target-present lineup, identifying the guilty suspect (i.e., the culprit) is the correct identification response, whereas identifying a foil or rejecting the lineup are incorrect responses. In a target-absent lineup, rejecting the lineup is the correct identification response, whereas identifying the innocent suspect or identifying a foil are incorrect responses. It is important to note, however, that in real life criminal investigations, only incorrect identifications of innocent suspects result in criminal proceedings being brought against that person, because incorrect identifications of foils are known errors.

Using these methods, researchers can examine the identification responses made by subjects. Often, this is achieved by computing the proportion of subjects who made each identification response. For instance, the correct identification rate (or, hit rate: HR) of guilty suspects in target-present lineups, is calculated by taking

the number of subjects who correctly identified the guilty suspect and dividing this by the number of target-present lineups. Similarly, the false identification rate (or, false alarm rate: FAR) of innocent suspects in target-absent lineups, is calculated by taking the number of subjects who incorrectly identified the innocent suspect and dividing this by the number of target-absent lineups. Let's say 100 subjects in a study saw a target-present lineup and 100 subjects saw a target-absent lineup. If 50 subjects correctly identified the guilty suspect and 25 subjects incorrectly identified the innocent suspect, then the correct identification rate (or HR) would be 50 ÷ 100 = .50, and the false identification rate (or FAR) would be 25 ÷ 100 = .25. Proportions of foil identifications and lineup rejections are calculated in a similar manner (see Table 1.1). That is, the incorrect identification rate of foils in target-present lineups is the number of subjects who incorrectly identified a foil from a target-present lineup (e.g., 30) divided by the number of target-present lineups (e.g., 100). The incorrect identification rate of foils in target-absent lineups is the number of subjects who incorrectly identified a foil from a target-absent lineup (e.g., 25) divided by the number of target-absent lineups (e.g., 100). Similarly, the incorrect rejection rate is the number of subjects who incorrectly rejected a target-present lineup (e.g., 20) divided by the number of target-present lineups (e.g., 100), while the correct rejection rate is the number of subjects who correctly rejected a target-absent lineup (e.g., 50) divided by the number of target-absent lineups (e.g., 100).

Table 1.1
Response Options in a Standard Eyewitness Identification Experiment

	Identification response		
Target presence	Suspect	Foil	Lineup rejection
Target present	Correct	Incorrect	Incorrect
	50 ÷ 100 = .50	30 ÷ 100 = .30	20 ÷ 100 = .20
	(hit rate; HR)		
Target absent	Incorrect	Incorrect	Correct
	25 ÷ 100 = .25	25 ÷ 100 = .25	50 ÷ 100 = .50
	(false alarm rate; FAR)		

Note. Hypothetical proportions of identification responses are calculated assuming that 100 target-present lineups and 100 target-absent lineups were presented.

The most frequently used accuracy measure in the eyewitness identification literature is the diagnosticity ratio, or the posterior of odds of guilt. The diagnosticity ratio focuses on suspect identifications to compute a single measure of performance, which is HR ÷ FAR (e.g., Steblay, Dysart, & Wells, 2011; Wells & Lindsay, 1980). If the HR is .50 and the FAR is .25, then the diagnosticity ratio is 2. A diagnosticity ratio of 2 implies that the suspect is twice as likely to be identified when guilty than when innocent. Another common measure is the posterior probability of guilt, which is HR ÷ (HR + FAR), with higher values indicating a higher probability that the suspect is guilty (Wells & Lindsay, 1980).

More than 40 years of research using these experimental methods have repeatedly shown that people make incorrect identification decisions (e.g., Cutler & Penrod, 1995; or see Clark, 2012 and Steblay, Dysart, Fulero, & Lindsay, 2001 for more recent meta-analyses). But, collectively, these studies show that the rate of errors vary greatly across studies, ranging from a few percent, to more than 90% (Wells, 1993). This shows that the accuracy of identifications is dependent on a host of different factors; some of these factors are not under the control of the Criminal Justice System, but some factors are. Those factors that are not under the control of the Criminal Justice System are called estimator variables, because their influence on identification accuracy in real cases can only be estimated post hoc (Wells, 1978). Research on estimator variables has improved knowledge about the elements of a criminal event that may reduce the likelihood that a witness makes a correct identification. Research shows that, for example, cross-race identifications are often less accurate than identifications of same race faces (Meissner & Brigham, 2001), stress can have an adverse effect on attention and memory (Deffenbacher, Bornstein, Penrod, & McGorty, 2004), and the presence of a weapon can impair identification performance (Fawcett, Russell, Peace, & Christie, 2013). System variables, on the other hand, are under the control of the Criminal Justice System and include factors such as police procedures and techniques for constructing lineups (Wells, 1978). Research in this regard has shown, for instance, that leaving the suspect to stand out because he looks different to the foils increases the number of suspect identifications (e.g., Clark, 2012; Doob & Kirshenbaum, 1973; Wells, Leippe, & Ostrom, 1979), while presenting lineup images one at a time (i.e., a sequential lineup) instead of together (i.e., a simultaneous lineup) reduces the number of suspect identifications

(Clark, 2012; R. C. L. Lindsay & Wells, 1985). Research suggesting that system variables could be modified to enhance identification accuracy can have direct implications on legal policy and procedures. Indeed, researchers have made recommendations for best practice to the Criminal Justice System (e.g., Brooks, 1983; Technical Working Group for Eyewitness Evidence, 1999; Wells et al., 1998). Although the benefits of some of these recommendations have recently been questioned (a point that I return to in the "Measurement issues" section; see Clark, 2012 and Gronlund, Mickes, Wixted, & Clark, 2015 for reviews), it is clear that real life lineup procedures have the propensity to be improved by experimental research. Accordingly, government agencies and policymakers around the world are calling for an increase in evidence-based practice (e.g., Cabinet Office, 2015; National Institute of Justice, 2016; Sherman, 1998). That is, there is an increased desire for procedures to derive from a solid base of scientific evidence about what works best.

Suspects with distinctive features

One procedure that is not currently evidence-based concerns how police officers construct lineups when the suspect has a distinctive facial feature (e.g., a tattoo, scar, piercing). Estimates of the number of suspects who have a distinctive feature are surprisingly high. Distinctive physical features were noted by police in the arrest report for over a third of defendants in San Diego, in the US (Flowe, Ebbesen, Libuser, Burke, & VanNess, 2010) and over one third of all police lineups in England and Wales contain a distinctive suspect (P. Burton, West Yorkshire Police, personal communication, November 3, 2008, as cited in Zarkadi, Wade, & Stewart, 2009). Despite this frequency, little research has examined the efficacy of the different methods that the police currently use when constructing lineups for distinctive suspects.

Current procedures

Lineups around the world typically contain the police suspect (who is either innocent or guilty) and a number of known innocent foils, but the way in which the lineup is presented to the witness varies across countries and jurisdictions. The most commonly used lineup procedure in the US, for instance, involves the simultaneous presentation of a number of photos—usually 6—taken of each person facing the camera. In England and Wales, however, the standard procedure involves the

sequential presentation of at least 9 video clips. In each video clip, the person first faces the camera, then moves their head to show their right profile, then to show their left profile, and then back to face the camera. Witnesses are shown the sequence of video clips twice before they are asked to make their identification decision, but witnesses can request to see the clips as many times as they wish (Horry, Memon, Milne, Wright, & Dalton, 2013; Police and Criminal Evidence Act 1984, Code D, 2011; see Seale-Carlisle & Mickes, 2016 for an empirical comparison of US and UK lineup procedures).

Although the US and the UK use different presentation methods, both countries—and a number of others, in fact—are guided by the same central principle when constructing lineups for suspects with distinctive features. Guidelines suggest that police officers must prevent distinctive suspects from standing out to ensure that every lineup member is a plausible alternative to the suspect (e.g., Brooks, 1983; Police and Criminal Evidence Act 1984, Code D, 2011; Technical Working Group for Eyewitness Evidence, 1999). To this end, guidelines state that police officers should either artificially replicate a suspect's distinctive feature across the lineup members (replication; see Figure 1.1b); or they should conceal the feature on the suspect and conceal a similar area on the other members (Police and Criminal Evidence Act 1984, Code D, 2011; Technical Working Group for Eyewitness Evidence, 1999). In practice, concealment usually involves either pixelating the area of the feature (pixelation; Figure 1.1c) or covering the area with a solid black rectangle (block; Figure 1.1d). The Police and Criminal Evidence Act in England and Wales (2011) further specifies that replication may be more appropriate when the witness has described the distinctive feature, whereas concealment may be more appropriate when the witness has not. Conversely, the Technical Working Group for Eyewitness Evidence (2003) in the US recommends that replication is the preferred technique, regardless of the witness's description. While these suggestions are provided, it is clear from the guidelines that the identification officer overseeing the case has discretion to choose whether to replicate or conceal the feature (Police and Criminal Evidence Act 1984, Code D, 2011).

Very little information exists on what police officers choose to do to deal with a distinctive feature on the face of a suspect. In the UK, there is some suggestion that concealment techniques may be used more frequently than replication techniques,

simply because concealment is usually cheaper, faster, requires less skill, and can be applied to moving video images, whereas replication techniques cannot (Horry et al., 2013; A. Monaghan, National VIPER User Group, personal communication, August 15, 2016). Some data on US practices exists, but the data were collected over 10 years ago. Wogalter, Malpass, and Mcquiston (2004) report the responses to a 67-item questionnaire that was completed by the most experienced lineup administrator in 220 different jurisdictions. One item asked what was done when the suspect has a distinctive feature, and then provided the officers with a list of non-mutually exclusive options. The majority of officers (77%) reported that they simply tried to select foils who had similar features, 23% and 18% reported using replication and concealment techniques, respectively. Perhaps somewhat surprisingly, given the legal guidelines, 30% of officers reported that they did not do anything to deal with a suspect's distinctive facial feature. Presumably, this means that the lineup was unfair because the distinctive suspect was left to stand out.

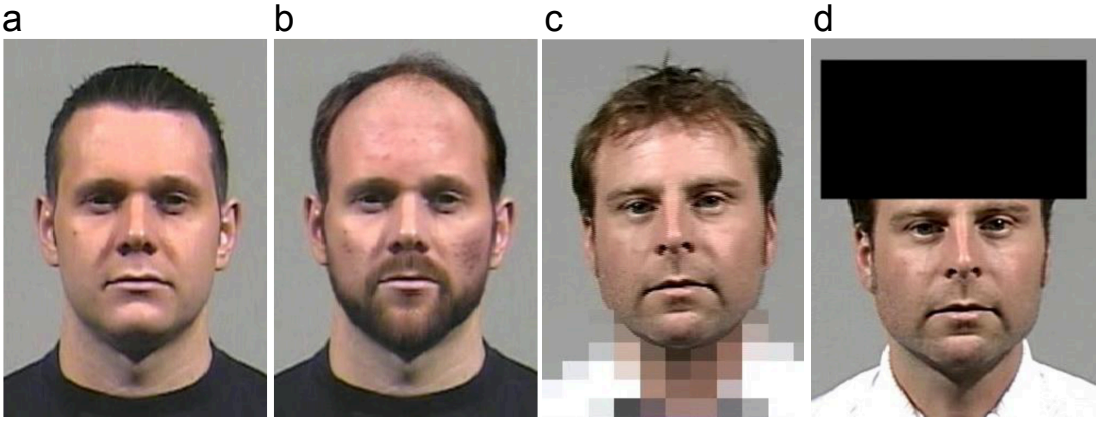

Figure 1.1. (a) An unaltered image, (b) an example of how features (baldness, facial hair and blemishes) can be digitally added for replication lineups, and examples of how features can be concealed using (c) pixelation, or (d) block techniques. Adapted from "Eyewitness Identification: Improving Police Lineups for Suspects with Distinctive Features," by T. Zarkadi, 2009, *Doctoral dissertation.* Images originally provided by the National Video Identification Parade Electronic Recording (VIPER) Bureau.

Research on lineups for distinctive suspects

But which lineup technique for distinctive suspects fosters the most accurate eyewitness identifications? Very little evidence exists in this regard, too. To date, only two published studies have examined the police techniques used to prevent suspects from standing out, and the results converge to suggest that replicating a

distinctive feature may enhance eyewitness identification performance more than removing it (Badham, Wade, Watts, Woods, & Maylor, 2013; Zarkadi et al., 2009). In these studies, subjects studied a set of greyscale images of faces, some of which had a distinctive feature. After a brief filler task, subjects attempted to recognise the distinctive faces they had previously studied from a series of lineups. In half of these lineups, the target's feature had been digitally added to the other lineup members. In the remaining lineups, the target's feature had been removed. Compared to removing the feature, replication increased correct identifications by approximately 20% in target-present lineups, without boosting incorrect identifications in target-absent lineups (Zarkadi et al., 2009). Although there seems to be little advantage of replicating distinctive features for older adults (aged 61–91), Badham et al. showed again that replication enhanced identification performance relative to removing the feature in younger adults (aged 18–24).

Together, the studies by Zarkadi et al. (2009) and Badham et al. (2013) addressed important theoretical and applied questions, but they do not provide information about how the lineup techniques compare to when nothing is done to prevent the distinctive suspect from standing out. Moreover, in practice, police officers do not remove distinctive features from suspects in lineups. It is possible that subjects made more incorrect rejections in target-present removal lineups than in replication lineups, because the person they believed to be the target was now missing a prominent distinctive feature that they remembered (Wixted & Mickes, 2014). In real criminal investigations, the feature is concealed—using pixelation or block techniques—which indicates that there could be a distinctive feature under the concealed area, and this may lead to a different pattern of identification responses. Finally, these studies calculated the proportion of identification responses (as demonstrated in Table 1.1), and new research suggests that this might not be the most appropriate way to measure identification accuracy when comparing different lineup techniques.

Measurement issues

Relatively recently, the best practice recommendations made to the Criminal Justice System (e.g., Technical Working Group for Eyewitness Evidence, 1999) and the research on which these were based have been criticised, because it has been suggested that proportion correct, diagnosticity ratios, or other closely related ratio-

based measures, should not be used to evaluate the effectiveness of different lineup techniques (Mickes, Flowe, & Wixted, 2012; Wixted, Gronlund, & Mickes, 2014). Critically, ratio measures cannot provide evidence that a particular procedure is superior, because they change systematically as a function of witnesses' willingness to make an identification decision—their response bias (Wixted & Mickes, 2014). Specifically, diagnosticity ratios increase as responding becomes more conservative (see Gronlund et al., 2012 and Mickes et al., 2012 for empirical demonstrations of this effect). A higher diagnosticity ratio, then, may simply reflect that a particular lineup procedure decreases both the FAR (a desirable outcome) and the HR (an undesirable outcome), compared to an alternative procedure. Yet, when two lineup procedures are compared, the better procedure is the one that decreases the FAR but also increases the HR, regardless of witnesses' willingness to choose. Or, to put it another way, the best lineup procedure is the one that best helps witnesses to discriminate between innocent and guilty suspects, regardless of how likely they are to choose from the lineup. Willingness to choose is under the control of the witness and can be easily varied over a wide range by making simple adjustments to procedures. Instructing witnesses that it is important that they identify the culprit even when their certainty is low, for instance, will make witnesses more likely to choose from the lineup, whereas instructing witnesses that it is important that they only make an identification when they are totally certain, will make them less likely to choose (Wixted & Mickes, 2014). As such, to adequately assess which lineup procedure is superior, one must measure discriminability—the ability to tell the difference between innocent and guilty suspects—separately from response bias. And to measure these two components, we can use signal detection theory.

Signal detection theory

Signal detection theory (SDT) describes how people make decisions in the presence of uncertainty (Green & Swets, 1966; Macmillan & Creelman, 2005). In a typical recognition memory test, subjects first study a set of items, say a list of words, and are subsequently presented with a test list that contains some words that were previously studied (targets) and some that were not (lures). According to SDT, the test items vary in memory strength, that is, some items feel more familiar than others. Targets that have been studied before feel, on average, more familiar (i.e., have a higher average memory strength) than lures that have not been studied before.

Displayed on a graph, the memory strength distribution for the targets lies higher along the memory strength axis than does the distribution for the lures, and the distributions are generally assumed to be Gaussian in form (see Figure 1.2). A subject's ability to tell the difference between target words and lure words is represented by the degree of overlap between the target and lure distributions. If there is a greater overlap, then this illustrates that a subject finds it more difficult to tell the difference between targets and lures. If there is little overlap, then this illustrates that a subject finds it easier to correctly sort targets and lures into their appropriate categories.

According to SDT, a decision criterion is placed on the memory strength axis. When the memory strength (or the feeling of familiarity) of a word exceeds this decision criterion, then the word is judged to be one that was previously studied (i.e., it is judged to be "old"). The HR is represented by the proportion of the target distribution that falls to the right of the decision criterion, whereas the FAR is the proportion of the lure distribution that falls to the right of the decision criterion (depicted by the light grey and dark grey shaded areas on Figure 1.2a, respectively). The placement of the decision criterion depends on a range of factors (environmental, experimental, internal), which increase or decrease how much information is required to accept a word as one that has been seen before. Imagine, for instance, that we add a financial reward to our word recognition study. For each word, all subjects can choose whether or not they make a decision. One group of subjects, let's call them the neutral group, receive £1 when they correctly identify a target word as old (i.e., make a hit), but also lose £1 when the incorrectly identify a lure word as old (i.e., make a false alarm). Another group of subjects, let's call them the conservative group, receive £1 for each hit, but they lose £10 when they make a false alarm. Because the cost of making an error is much higher in the conservative group, they will respond to fewer items. They are likely to only state that an item is old when they are very likely to be correct (i.e., when the feeling of familiarity is very high). Theoretically, they have set their decision criteria higher on the memory strength axis; they have set a more conservative decision criterion (Figure 1.2b) than subjects in the neutral group (Figure 1.2a). Setting a more conservative decision criterion produces fewer hits and false alarms, as shown by the smaller proportion of the target and lure distributions that fall to the right of the decision criterion in Figure

1.2b. Perhaps we introduce a third group of subjects, the liberal group, in which the pay out for a hit is £10, but subjects lose £1 if they make a false alarm. In this case, the benefit of getting a hit is much greater, so subjects would set a more liberal criterion (Figure 1.2c), which produces a greater number of both hits and false alarms. The key point is that in each group, subjects' ability to tell the difference between targets and lures—the distance between the two memory strength distributions—is the same, but changes in the decision criterion across the groups can lead to very different HRs and FARs. When we conceptualise a decision task using SDT, then, it is clear that performance is determined by both ability to tell the difference between targets and lures (the overlap of the distributions), and response bias (the placement of the decision criterion). Thus, both of these distinct elements need to be measured if we are to understand memory performance in different lineup techniques.

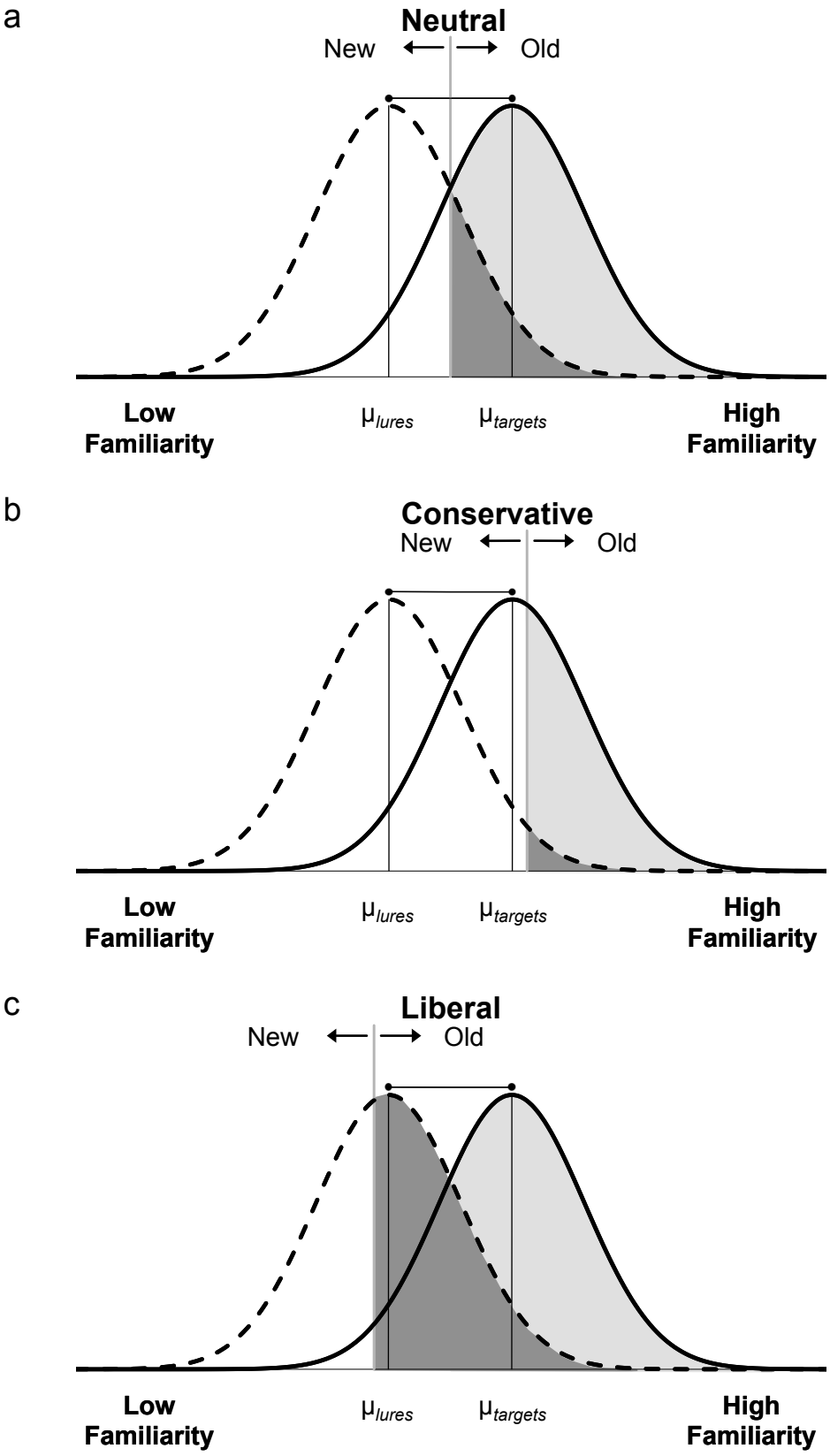

Figure 1.2. Signal detection model with (a) a neutral, (b) a conservative, and (c) a liberal decision criterion. The dashed distribution represents the memory strength of lures and the solid distribution represents the memory strength of targets. The proportion of hits and false alarms are represented by the light grey and dark grey shaded areas, respectively.

Indeed, SDT can be applied to eyewitness decision-making (e.g., Clark, 2003; Wixted & Mickes, 2014). The SDT model describes the distribution of memory strengths associated with guilty suspects, innocent suspects and foils for a group of subjects tested under a particular set of conditions; a group of subjects viewing a particular lineup procedure, for instance (Wixted & Mickes, 2014). When a witness views the faces in a lineup, each face has some memory strength value (i.e., degree of familiarity). Guilty suspects, innocent suspects and foils each have memory strength values with Gaussian distributions and means of μ_{guilty}, $\mu_{innocent}$, and μ_{foil}, respectively. In a fair lineup, in which all of the lineup members are plausible alternatives to the culprit, the innocent suspect is not more similar to the guilty suspect than the other foils, so $\mu_{innocent} = \mu_{foil}$. Therefore, the model for a fair lineup consists of two distributions: one for guilty suspects (μ_{guilty}), and one for innocent suspects and foils ($\mu_{innocent}$; see Figure 1.3a). The guilty suspect distribution is simply the target distribution from our word memory experiment, and the innocent suspect and foil distribution is the lure distribution. The guilty suspect distribution (μ_{guilty}) is situated higher on the decision axis than the distribution for innocent suspects and foils ($\mu_{innocent}$), which reflects the idea that, on average, guilty suspects are associated with a greater memory strength (i.e., feel more familiar) than innocent suspects and foils who have not been seen before.

The model differs slightly for an unfair lineup, in which the innocent suspect is more similar to the guilty suspect than the other foils. In this case, the model consists of three distributions: one for guilty suspects (μ_{guilty}), one for innocent suspects ($\mu_{innocent}$), and one for foils (μ_{foil}, see Figure 1.3b). In an unfair lineup, then, there are three different discriminabilities that can be measured: the ability to discriminate (a) guilty suspects from innocent suspects, (b) guilty suspects from foils, and (c) innocent suspects from foils; and each is measured by the degree of overlap between the two distributions being considered. Yet, from a practical stand-point, in both fair and unfair lineups, subjects' ability to discriminate guilty suspects from innocent suspects (i.e., the distance between the μ_{guilty} and $\mu_{innocent}$ distributions) is the key discriminability to measure, because identifications of foils do not result in any legal action against the foil that is selected. Again, greater overlap of the μ_{guilty} and $\mu_{innocent}$ distributions reflects poorer ability to tell the difference between guilty and innocent suspects.

As before, a decision criterion is placed on the memory strength axis. When a face is familiar enough to exceed the decision criterion (denoted as c_1 on Figure 1.3a and Figure 1.3b), then a positive identification is made. The simplest decision rule (but not the only possible decision rule, see Clark, Erickson, & Breneman, 2011; Fife, Perry, & Gronlund, 2014) is that an eyewitness determines which lineup member best matches their memory of the culprit, and then they identify this face if its familiarity value exceeds c_1. If no face is familiar enough to exceed c_1, then the witness states that the real culprit is not in the lineup (i.e., they make a lineup rejection). In both fair and unfair lineups (and other decision-making tasks) more decision criteria can be added to the SDT model (c_2, c_3, c_4, c_5). The criteria higher on the decision axis represent identification decisions that are made with greater memory strength. They are more conservative identification decisions; identifications that are made with greater levels of confidence. A witness decides to identify a face with low levels of confidence (perhaps she is "10% certain") when, theoretically, the memory strength of the face exceeds c_1, but not c_2. A witness decides to identify a face with the next level of confidence (perhaps she is "30% certain"), when the memory strength of the face exceeds c_2, but not c_3, and so forth. The decision to identify a face with the highest level of confidence (i.e., when she is "100% certain") means that the memory strength is strong enough to exceed the highest criterion (here, c_5). In SDT, a positive identification decision and a confidence rating are theoretically equivalent, because both are based on a decision criterion placed on the decision axis (Macmillan & Creelman, 2005; Wixted & Mickes, 2014). Adding more decision criteria to the SDT model, by asking subjects to provide a confidence rating, is therefore a way to gain more information about subjects' recognition memory.

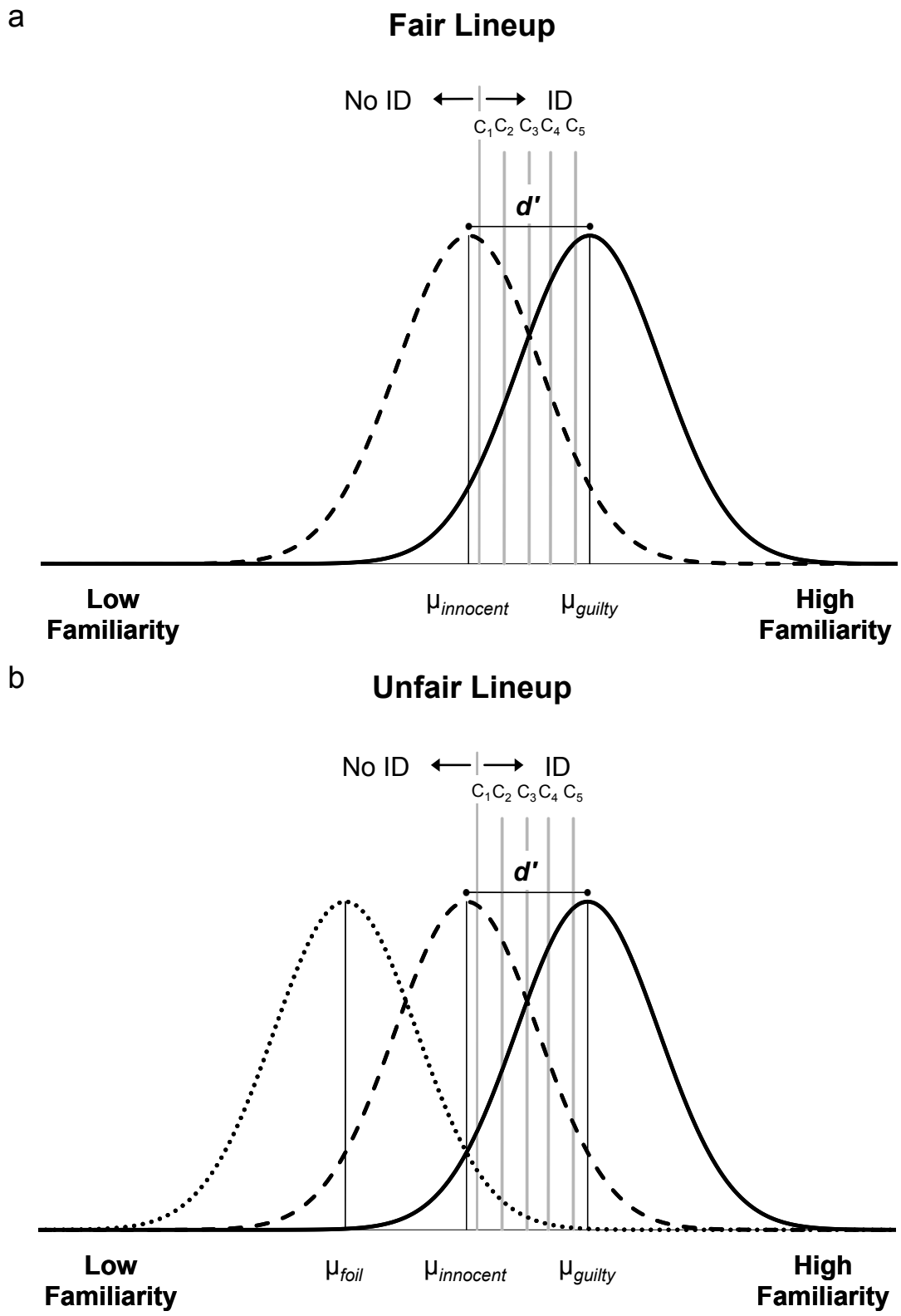

Figure 1.3. Signal detection model for (a) a fair lineup and (b) an unfair lineup (Wixted & Mickes, 2014). In a fair lineup, the dashed distribution represents the memory strength of innocent suspects and foils. In an unfair lineup, the dashed distribution represents the memory strength of innocent suspects, and the dotted distribution represents the memory strength of foils. In both fair and unfair lineups, the solid distribution represents the memory strength of guilty suspects. c_1, c_2, c_3, c_4 and c_5 are a set of response criteria that reflect different levels of confidence.

Measuring discriminability

Policymakers in the Criminal Justice System should seek to employ lineup procedures that enhance witnesses' ability to discriminate between innocent and guilty suspects. How should researchers measure this element of performance? Two measures of discriminability are d' and receiver operating characteristic (ROC) analyses. Looking back to Figure 1.3, d' measures the distance between μ_{guilty} and $\mu_{innocent}$ in standard deviation units. d' is calculated by transforming the HR and FAR to z scores, which converts the HR and FAR to standard deviation units, and then taking the difference, $d' = z(\text{HR}) - z(\text{FAR})$. When the two distributions overlap completely, $d' = 0$. Thus, higher values of d' indicate less overlap of the μ_{guilty} and $\mu_{innocent}$ distributions and therefore reflect better discriminability. However, d' estimates performance using one HR-FAR pair and some theoretical assumptions (i.e. the assumptions of SDT displayed in Figures 1.2 and 1.3). A better method to characterise identification accuracy is to use ROC analyses, because this is a theory-free technique that does not make any assumptions about the underlying distributions of the data (Mickes, Moreland, Clark, & Wixted, 2014).

In ROC analysis, the first step is to construct an ROC curve for each lineup technique. Each curve plots the correct identification rate of guilty suspects in target-present lineups (hit rate; HR) against the false identification rate of innocent suspects in target-absent lineups (false alarm rate; FAR). In many ways, ROC analysis is like the traditional diagnosticity ratio, determined by HR ÷ FAR (e.g., Steblay et al., 2011). But instead of calculating a single diagnosticity ratio (one HR-FAR pair), we plot several HR-FAR pairs over decreasing levels of confidence. Confidence serves as a proxy for willingness to choose, with decreasing levels of confidence equating to more liberal responding (Wixted & Mickes, 2014). Therefore, by plotting these HR-FAR pairs over the full range of confidence, we can determine how the different lineup types affect subjects' ability to distinguish between real culprits and innocent suspects, independently of their willingness to identify the suspect (Gronlund, Wixted, & Mickes, 2014; National Research Council, 2014).

Figure 1.4 displays this thinking more concretely, and depicts two hypothetical ROC curves. Let's say subjects made their confidence rating on an 11-point Likert-type scale (0% = c*ompletely uncertain* to 100% = *completely certain*). The lowest left point of each curve, highlighted in grey, represents the HR and FAR at the

highest level of confidence ("100% certain"). The second point on each curve represents the HR and FAR at the highest level of confidence and the second highest level of confidence (i.e. "100% certain" and "90% certain"), and so forth. As one moves along the curve, one eventually reaches the farthest right point (circled in grey), which shows the rates for all subjects who made an identification. A key idea is that for any point on the lower ROC (white circles), there is an achievable point on the higher ROC (solid black circles) that is associated with both a higher HR and a lower FAR. Therefore, the ROC curve that falls closest to the upper left corner of the plot—closest to the star and farthest from the dashed chance line—is the objectively superior procedure because it maximises guilty suspect (i.e., culprit) identifications while minimising innocent suspect identifications. Put simply, this procedure allows witnesses to most accurately discriminate between guilty and innocent suspects.

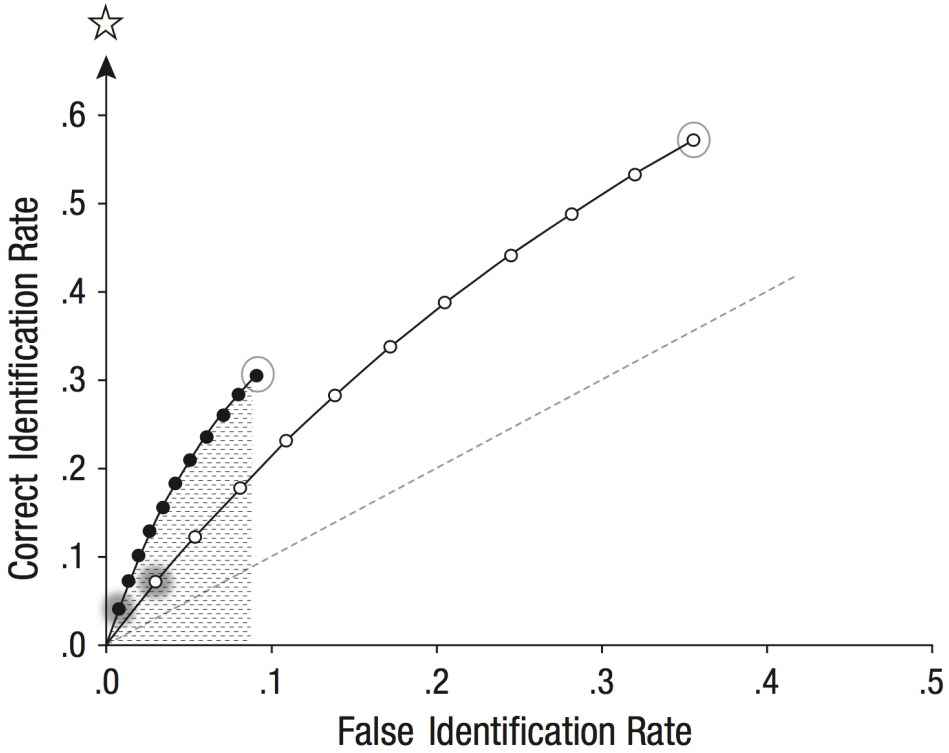

Figure 1.4. Two hypothetical receiver operating characteristic (ROC) curves. The curve through the black operating points is the lineup procedure that allows witnesses to most accurately discriminate between guilty and innocent suspects, because it falls closest to perfect accuracy (the star, where hit rate = 1 and false alarm rate = 0) and farthest from the dashed chance line compared to the alternative procedure. The lowest left point on each curve (highlighted in grey) represents the correct and false suspect identifications made with the highest level of confidence, whereas the farthest right point (circled in grey) represents the rates for all subjects who made an identification. The partial Area Under the Curve (*p*AUC) for the shaded area under the curve with the solid black circles is calculated by setting the specificity (1 – false alarm rate at the right-most edge of the shaded area) to .91.

To compare ROC curves, we compare the partial Area Under the Curve (*p*AUC) because the FAR for innocent suspects is less than 1. In *p*AUC analysis, one defines the specificity (1 – FAR) for calculating the AUC. For example, if we were interested in the calculating the shaded area under the curve with the solid black circles in Figure 1.4, we would calculate the *p*AUC statistics by defining the specificity as (1 – .09) = .91. When comparing ROC curves, the specificity must be set to the same value in every *p*AUC calculation. Thus, in the current example, when calculating the area under the curve with the white circles, we would also set the specificity as .91. The ROC curve that produces the largest *p*AUC is the procedure that best enables witnesses to discriminate between guilty and innocent suspects.

Despite some criticism (e.g., Lampinen, 2016; Wells, Smalarz, & Smith, 2015), a recent National Academy of Sciences report endorsed the notion that policymakers should seek to employ lineup procedures that best enable witnesses to discriminate between guilty and innocent suspects and recommended ROC analyses, over ratio-based measures, for that task (National Research Council, 2014). Although ratio-based measures do not provide the information needed by policymakers, they do help to provide the information needed by judges and jurors—that is, the likely accuracy of an identification made with a particular level of confidence.

Gauging the likely accuracy of identifications

Regardless of what procedure has been used to collect the identification evidence, judges and jurors want to know the likely accuracy of an identification made with a particular level of confidence, because this provides them with information about whether an identification is likely to be reliable (Juslin, Olsson, & Winman, 1996; Mickes, 2015). With this in mind, much research has set out to examine whether witnesses are able to assess the likely accuracy of their memories and assign appropriate confidence judgements. Do witnesses express high confidence in their decision when their answer is correct, and lower confidence when their answer is incorrect? A vast body of research has assessed the relationship between confidence and accuracy by calculating the correlation coefficient and, although the relationship is stronger amongst witnesses who choose someone from the lineup (e.g., Sporer, Penrod, Read, & Cutler, 1995), the general conclusion has been that eyewitness confidence is not a reliable indicator of accuracy (e.g.,

Bothwell, Deffenbacher, & Brigham, 1987; Lacy & Stark, 2013; Penrod & Cutler, 1995).

But we now know that a low correlation coefficient does not necessarily indicate a poor relationship between confidence and accuracy (Juslin et al., 1996). Correlation coefficients reflect the relationship between categorical confidence judgements (0, 10, 20, etc.) and binary accuracy (correct or incorrect). When displayed in a graph, confidence is plotted on the x-axis and accuracy (correct or incorrect) on the y-axis, and each point represents the confidence and accuracy of one person. Computing the correlation coefficient involves fitting a straight line through these data, and the distribution of confidence judgements heavily influences the line. Confidence judgements made by subjects in empirical studies are usually made within a relatively restricted range (i.e., the distribution of confidence judgements is unimodal) and this serves to underestimate the relationship between confidence and accuracy (Juslin et al., 1996; D. S. Lindsay, Read, & Sharma, 1998). Furthermore, because accuracy is plotted as a binary outcome for each person, correlation coefficients do not provide information about the likely accuracy of an identification made with a particular level of confidence (Brewer & Wells, 2006; Juslin et al., 1996). A more suitable statistical technique for testing whether people can assess the likely accuracy of their memories is (to use ratio-based measures) to plot *average* accuracy at different levels of confidence—that is, plot confidence-accuracy curves. Only this technique tells us the likely accuracy of an identification made with a particular level of confidence. It also remains unaffected by the distribution of confidence judgements because average accuracy (i.e., probability of a correct identification decision) at a particular level of confidence is the same, regardless of the number of identifications made at that level of confidence (Brewer, Keast, & Rishworth, 2002; Brewer & Wells, 2006; Juslin et el., 1996; Mickes, 2015; Wixted & Wells, 2016).

Studies that plot confidence-accuracy curves now show that the confidence judgement taken at the time of the identification decision is often meaningfully related to likely accuracy: a finding that has been seen in both the lab (e.g., Brewer & Wells, 2006; Horry, Palmer, & Brewer, 2012; Mickes, 2015, Experiment 1; Palmer, Brewer, Weber, & Nagesh, 2013; Sauer, Brewer, Zweck, & Weber, 2010; Weber & Brewer, 2004; Wixted, Mickes, Clark, Gronlund, Roediger, 2015; Wixted,

Read, & Lindsay, 2016) and the field (Sauerland & Sporer, 2009; Wixted, Mickes, Dunn, Clark, & Wells, 2016; see Wixted & Wells, 2016 for a review). There are some instances, however, in which confidence is uninformative of accuracy (e.g., Chandler, 1994; Mickes, 2015, Experiment 2; Sampaio & Brewer, 2009; Shaw & McClure, 1996; Wells & Bradfield, 1999). Roediger, Wixted, and DeSoto (2012) succinctly summed up the relation between confidence and accuracy with: "it depends" (p. 85). The authors stated: "…eyewitness memory confidence is a useful but imperfect indicator of the truth." (p. 113, Roediger et al., 2012). Thus, plotting confidence-accuracy curves can help to further our understanding about the situations in which it may be appropriate for judges and jurors to use a witness's confidence statement as a proxy for their likely accuracy (Palmer et al., 2013).

In sum, although ratio-based measures of performance are useful for examining average accuracy at difference levels of confidence, they are not so useful when comparing the efficacy of different lineup procedures. An argument has been put forth as to why it is important to measure discriminability independently of response bias when assessing different lineup procedures and two measures of discriminability have been outlined. Now we need a theory of eyewitness discriminability to help interpret our findings.

Theory

One theory that has been dominant in the eyewitness identification literature is the distinction between relative and absolute judgements (Wells, 1984, 1993). A relative judgement is the tendency to choose the lineup member who looks most like the witness's memory of the culprit relative to the other lineup members, whereas an absolute judgement is the tendency to choose the lineup member whose match to the witness's memory of the culprit is sufficiently high, above some absolute criterion (Wells, 1984). But given that not everyone who uses a relative judgement strategy chooses someone from the lineup, they must also set a decision criterion (Ebbeson & Flowe, 2002). Therefore, the relative-absolute distinction can be conceptualised as a theory of response bias, with relative judgements reflecting a more liberal bias, and absolute judgements reflecting a more conservative bias (Wixted & Mickes, 2014). Moreover, the relative-absolute distinction is a verbally specified theory, but formally specified mathematical models are arguably more beneficial for theory development because they are more precise and readily falsifiable (Clark, 2008). As

such, a formally specified model of eyewitness discriminability has recently been proposed—the diagnostic-feature-detection model (Wixted & Mickes, 2014).

Diagnostic-feature-detection model

The diagnostic-feature-detection model starts with a familiar premise of SDT: for each lineup member's face, the features combine to create a memory signal (i.e., a feeling of familiarity), and the witness uses that signal to make their identification decision (Wixted & Mickes, 2014). The model suggests that some facial features differ between innocent and guilty suspects and are therefore diagnostic of guilt, whereas other facial features are shared by innocent and guilty suspects and are therefore non-diagnostic. The non-diagnostic features are those that correspond to the description of the culprit provided by the eyewitness. Whether innocent or guilty, the suspect will have those features, which means that relying on those features to decide whether or not the culprit is in the lineup will harm performance. The key premise of the model, then, is that witnesses are better at discriminating between innocent and guilty suspects when they base their decisions on (diagnostic) facial features that differ between innocent and guilty suspects, rather than on (non-diagnostic) facial features that innocent and guilty suspects share. It follows that identification procedures that best enhance witnesses' ability to discriminate between innocent and guilty suspects are those procedures that make it clearest to witnesses that certain facial features are shared by all members and are therefore not useful in making the identification. Or, put another way, the identification procedures that best enhance witnesses' ability to discriminate between innocent and guilty suspects are those procedures that best accentuate the non-diagnostic features. This is because witnesses can then discount the non-diagnostic features, and, instead, rely on diagnostic features that are unique to the guilty suspect.

Support for the diagnostic-feature-detection account comes from its ability to explain why simultaneous lineups, in which all of the faces are presented together, enhance subjects' ability to tell the difference between innocent and guilty suspects more than sequential lineups, in which the faces are presented one at a time (Dobolyi & Dodson, 2013; Gronlund et al., 2012; Mickes et al., 2012) and showups, in which a single image of the suspect is presented (Clark, 2012; Gronlund et al., 2012; Key et al., 2015; Neuschatz et al., 2016; Wetmore et al., 2015). In fair simultaneous lineups the foils and the suspect all match the description of the culprit, so, according to the

diagnostic-feature-detection account, presenting their photos simultaneously accentuates the non-diagnostic features. By contrast, presenting a face on its own in a showup or as part of a sequential lineup does not easily permit comparison across multiple faces and therefore reduces the witness's opportunity to learn which facial features are shared. Witnesses may therefore rely to a greater extent on non-diagnostic features, which will impair their ability to discriminate between innocent and guilty suspects. Indeed, some research has shown that presenting suspects late in a sequential lineup can enhance discriminability more than presenting suspects early in the lineup (Carlson, Gronlund, & Clark, 2008; Gronlund, Carlson, Dailey, & Goodsell, 2009; Gronlund et al., 2012). Presumably, when more faces are presented before the suspect, this provides subjects with greater opportunity to observe shared features (Goodsell, Gronlund, & Carlson, 2010).

Further support for this theoretical account comes from a study in which the police suspect was left to stand out in a lineup. Witnesses were better able to discriminate between innocent and guilty suspects when all lineup members, including the suspect, had the same emotional expression compared to lineups in which only the suspect had that expression (Flowe, Klatt, & Colloff, 2014). Presumably, when all the lineup members shared the same expression, people discounted the expression, and used other, diagnostic, cues to make an identification. Conversely, when only the suspect had the expression, people used this to make their identification decision, which impaired discriminability, because the expression was something that both the innocent and guilty suspect shared.

Clearly, the findings of multiple studies can be explained by the diagnostic-feature-detection account, but, by and large, the diagnostic-feature-detection model was developed to account for these findings. The next step in theory refinement, is to make predictions based on the theoretical account and then test these predictions empirically. Fittingly, the diagnostic-feature-detection model makes clear predictions about how the lineup techniques—replication, pixelation, block—for distinctive suspects are likely to affect witnesses' ability to discriminate between innocent and guilty suspects. Therefore, not only does the diagnostic-feature-detection model provide an appropriate theoretical framework to conceptualise and interpret our data, but the research presented in this thesis also provides the first direct test of this new theory. Testing theoretical models of eyewitness decision-making is important,

because, once refined, theories can be used to develop procedures that enhance eyewitness identification accuracy in real world criminal investigations (Gronlund et al., 2015; Wixted & Mickes, 2014).

Thesis aims and outline

The main aims of this thesis are two-fold:

1. Investigate how the lineup techniques for distinctive suspects influence eyewitness identification performance, to help further our understanding of which lineup techniques for distinctive suspects may be most appropriate in real criminal investigations.
2. Test the diagnostic-feature-detection model, to help further our theoretical understanding of how eyewitnesses make identification decisions.

Chapter 2 examines how replication, pixelation, and block lineups influence identification performance compared to unfair "do-nothing" lineups, in which nothing was done to stop the distinctive suspect from standing out. Chapter 3 examines how replication, pixelation, block and do-nothing lineups influence identification performance in young, middle-aged and older adults and examines how identification performance changes with age. Chapter 4 examines how replication, pixelation, block and do-nothing lineups influence identification performance when the culprit does not have a distinctive feature during the crime, and compares this to performance on the same lineups when the culprit does have a feature during the crime. Chapter 5 focuses on the replication technique and examines, in two experiments, how variation in the way the suspect's feature is replicated across the foils influences identification performance. Finally, Chapter 6 presents a general discussion of these four chapters in the context of the eyewitness decision-making literature and discusses potential limitations and possible fruitful areas of further research.

Chapter 2 :
Identification Performance on Fair and Unfair Lineups

"In the instant case, there is absolutely no evidence of undue suggestion created by the procedures used…[The] defendant's photograph did not stand out from the rest..."

People v. Bethea (1971)

Overview

Eyewitness identification studies have focused on the idea that unfair lineups, in which the suspect stands out, make witnesses more willing to identify that suspect. We asked whether unfair lineups—featuring suspects with distinctive features—also influence subjects' ability to distinguish between innocent and guilty suspects, and their ability to judge the accuracy of their identification. In a single experiment (N = 8,925), we compared three fair lineup techniques used by the police to unfair lineups in which we did nothing to prevent distinctive suspects from standing out.

Introduction

In 1986, a woman viewed a lineup and identified Leonard Callace as her attacker. She had described the attacker as a White male with reddish-blonde, afro-style hair and a full beard. But Callace—who had a full beard, and straight hair—appeared in the lineup with five men who had only moustaches. After Callace served six years in prison, DNA evidence revealed he was not the attacker. Callace's case, and many others, highlights the importance of preventing suspects with distinctive features from standing out in lineups (see http://www.innocenceproject.org/). But why do unfair lineups impair eyewitness identification performance? Is it because unfair lineups make witnesses more willing to identify the suspect? Or is it because unfair lineups make it more difficult for witnesses to determine if the lineup contains the actual culprit? We aimed to answer these questions.

We know that suspects who stand out are prone to be selected for the wrong reasons—namely, not because they match the witness's memory of the culprit (Wells, Rydell, & Seelau, 1993). Why? The long-standing explanation is that witnesses tend to select the person who looks most like the culprit, much like the way a student answering a multiple choice question tends to select the option that

looks most like the right answer (Wells, 1984). Indeed, it is well established that when the only person who matches the witness's description of the culprit is the suspect, the witness tends to select the suspect instead of another lineup member (Doob & Kirshenbaum, 1973; Wells, Leippe, & Ostrom, 1979). More recent reviews and meta-analyses also show that when the suspect looks less like the other members of a lineup, witnesses identify the suspect more often (Clark, 2012; Fitzgerald, Price, Oriet, & Charman, 2013). Two problems arise from this tendency. First, if the suspect is the culprit (i.e., the suspect is guilty), the identification is correct, but not for the right reasons—much like the student who gets the correct answer but does not actually know the right answer. Second, if the suspect is not the culprit (i.e., the suspect is innocent), the misidentification might send an innocent person to prison. The observation that witnesses are more willing to identify the suspect—which means correctly identifying a guilty suspect when he is present in the lineup, but incorrectly identifying an innocent suspect when the real culprit is not present—can help us to understand why unfair lineups often result in misidentifications.

Yet, a new approach, the diagnostic-feature-detection model, supports an additional prediction: Unfair lineups may also impair witnesses' ability to differentiate between the actual culprit and an innocent suspect (Wixted & Mickes, 2014). To see why, consider what happens when a witness views the members in a lineup, whether fair or unfair. The idea is that for each lineup member's face, features combine to create a memory signal (a sense of familiarity and recollection) and the witness uses that signal to make an identification decision. Because some features differ between the culprit and an innocent suspect, they can help the witness make a better decision. For instance, Leonard Callace had straight hair, while the culprit had an afro. But other facial features are shared by the culprit and an innocent suspect, so they cannot help the witness. For instance, Callace and the culprit each had a full beard. If witnesses give weight to these shared features, their ability to distinguish between culprits and innocent suspects will suffer.

How, then, do witnesses make identifications in an unfair lineup, where only the suspect possesses the distinctive facial feature—say, a full beard—that the witnesses remember? To the extent witnesses do not realise the distinctive feature is unhelpful, they might erroneously weight that feature. Giving weight to an unhelpful feature will impair their ability to discriminate between real culprits and innocent

suspects. Consistent with this idea, one study showed that witnesses were better able to distinguish between guilty and innocent suspects when all lineup members, including the suspect, had the same emotional expression. But witnesses found it harder to distinguish between innocent and guilty suspects when the suspect was the only one with that expression (Flowe et al., 2014). Presumably, those subjects who saw the "matched expression" lineup discounted the shared emotional expression and used other, useful information to make an identification. By contrast, those who saw the "unmatched expression" lineup weighted the shared emotional expression, even though it was objectively unhelpful because it was something that both the innocent and guilty suspect shared. Other studies have found that people are better able to distinguish between innocent and guilty suspects when they are presented with a fair lineup rather than a single photo of a suspect (i.e., a showup, Key et al., 2015; Wetmore et al., 2015). Again, the fair lineup may permit subjects to discount unhelpful features but a single photo may not.

In the real world, police guidelines for constructing lineups often state that the police should prevent suspects with distinctive features from unduly standing out. In the US, England and Wales, for instance, police sometimes artificially replicate a suspect's distinctive feature across the lineup members (replication, see Figure 2.1a); other times, they conceal the feature on the suspect and conceal a similar area on the other members (Police and Criminal Evidence Act 1984, Code D, 2011; Technical Working Group for Eyewitness Evidence, 1999). Concealing involves either pixelating the area of the feature (pixelation, Figure 2.1b), or covering the area with a solid black rectangle (block, Figure 2.1c). These techniques represent a heartening translation of science into practice. Nonetheless, many efforts to make lineups fair are unsuccessful, and police officers still often do nothing and leave suspects to stand out (e.g. MacLin, MacLin, & Albrechtsen, 2006; Valentine & Heaton, 1999; Wogalter et al., 2004).

How, then, might replication, pixelation or block lineups affect eyewitness identification performance? First, because the suspect does not unduly stand out, witnesses should be less willing to identify the suspect. Second, because the distinctive feature appears either on every lineup member (replication), or on none of the lineup members (pixelation, block), witnesses should be more likely to weight something other than the distinctive feature. Therefore, they should also be better

able to distinguish between the culprit and an innocent suspect. By contrast, if a suspect is left to stand out (do-nothing lineups, Figure 2.1d), witnesses should be more willing to choose the suspect, and they should find it harder to distinguish between the culprit and an innocent suspect. The current research tested these hypotheses.

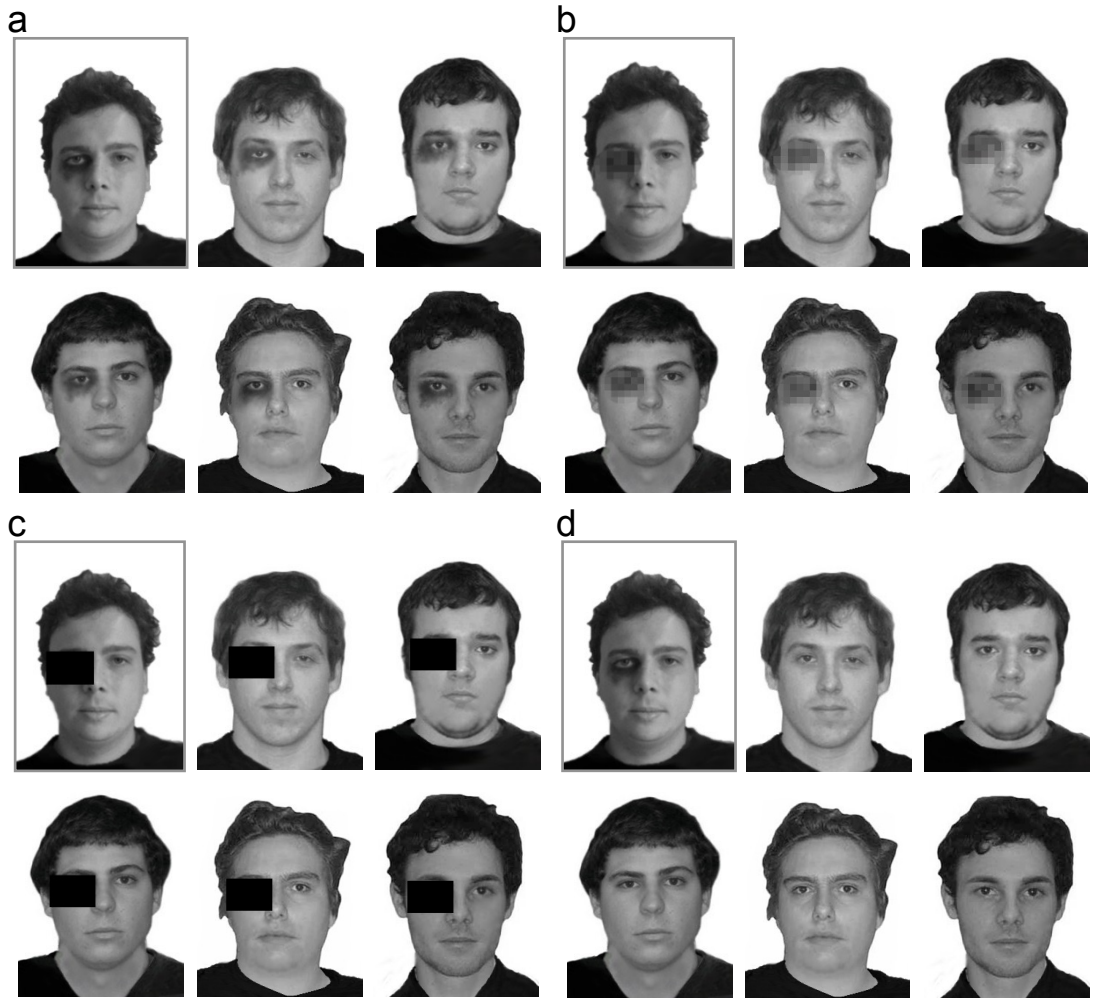

Figure 2.1. Examples of (a) a replication lineup, (b) a pixelation lineup, (c) a block lineup, and (d) a do-nothing (unfair) lineup. Top left image in each lineup is the suspect with the distinctive facial feature.

Method

Design

We used a 4 (lineup type: replication, pixelation, block, do-nothing) × 2 (target: present, absent) between-subjects design. Our data-collection stopping rule was to recruit as many subjects as possible before the end of spring term, with a minimum of 1,000 subjects with useable data in each of the eight conditions.

Subjects

The subjects were 9,841 adults from around the world who completed the task online. We excluded 916 people (10% in total; between 89–218 in each of the eight conditions), which resulted in a total sample size of 8,925. We excluded subjects who experienced technical difficulties while watching the video (*n* = 689, 7% in total), experienced programming errors while viewing the lineup (*n* = 128, 1% in total), or incorrectly answered an attention-check question on the content of the video (*n* = 99, 1% in total). The final sample consisted of 5,495 subjects recruited from social-networking sites who were entered into a prize draw for four £50 Amazon vouchers; 2,405 subjects recruited via Amazon Mechanical Turk who received $0.60; 871 students recruited from John Jay College of Criminal Justice who received extra credit in a course; and 154 students recruited from a sixth form (final year of high school) in the UK who completed the study as part of a research-methods course. Because the pattern of results was the same among the Internet and student samples, we combined the data for our analyses. Each cell contained between 1,017 and 1,145 subjects. We also checked for multiple responses by the same individual by examining IP addresses and e-mail addresses. These checks revealed 26 possible cases of duplicates (i.e., 0.003 of subjects). Our results are the same regardless of whether we include or exclude these people. Table 2.1 shows a demographic breakdown of the sample.

Table 2.1
Demographic Information For Social Media, Mechanical Turk, University, and Sixth Form Samples

	Social media	Mechanical Turk	University	Sixth form
Sex				
Male	1,498	1,091	265	40
Female	3,960	1,309	599	114
Prefer not to say	37	5	7	0
Age (years)				
16–20	1,606	79	593	149
21–30	1,693	997	252	0
31–40	870	675	18	0
41–50	649	326	4	0
51–60	395	224	0	0
61–70	161	86	0	0
≥ 71	46	13	0	0
Prefer not to say	75	5	4	5
Race or ethnicity				
White or European	4,633	1,494	195	72
Latin or Hispanic	52	102	339	0
Black, African, or Caribbean	72	178	140	6
South Asian	156	399	41	5
East Asian	175	90	42	6
Middle Eastern	25	7	13	2
Mixed	136	71	37	11
Other	147	41	27	39
Prefer not to say	99	23	37	13

Materials

Videos

It is widely documented that variability in encoding and test conditions is crucial when trying to detect reliable and generalizable effects (Brewer, Keast, & Sauer, 2010; D. S. Lindsay et al., 1998). Accordingly, we created four 30-s, non-violent videos depicting four different crimes, so that encoding conditions varied on several dimensions, including (a) the appearance of the target (each video featured a different, White, male culprit); (b) the distinctive feature on the target (each culprit

had a unique distinctive feature); (c) the crime committed (carjacking, graffiti, mugging, theft), and (d) the exposure duration of the target in each video (which ranged from 5 to 16 s across the four videos). At test, variation occurred between the encoding stimuli (the target in the crime video) and the test stimuli (the target's photographic image), simply because videos and photographs of people can vary to different extents. Targets also varied in their similarity to the foils.

In the *carjacking* scenario, a White female in her late 20s walks to her car, places her bag on the front passenger seat and sits in the driving seat preparing to drive off. A White male culprit in his mid-20s, with a large scar on his left cheek, opens the driver's door and instructs the female to get out of the car. The male gets in the car, rummages in the female's bag, and then starts the engine to drive off. In the *graffiti* scenario, a White male culprit in his early-20s, with severe bruising around his right eye, walks up to a wall, shaking a can of spray paint. After checking for witnesses, he uses the spray paint to write "UNI SUCKS" on the wall. In the *theft* scenario, a White male culprit in his early-20s, with a number of small nose piercings in his left nostril, walks down a university corridor. He enters an unlocked office and, after rummaging in a number of drawers, steals a laptop from a desk. In the *mugging* scenario, a White male in his late-20s is talking on his phone. A White male culprit in his early-20s, with a facial tattoo on his right cheek, approaches and instructs the victim to give him his phone. The victim refuses, but the culprit pushes him, snatches his phone, and runs off.

Lineups

We used 6-person simultaneous lineups that either contained the culprit and five foils (a target-present lineup), or contained six foils (a target-absent lineup). We created a pool of 40 foils for each culprit, so that we could randomly generate lineups from these pools. To create the pools of foils, we first asked a group of 18 subjects to watch each crime video and then answer 16 questions about the culprit's physical attributes, including questions about his gender, eye colour, hair colour, height, weight and ethnicity. Some characteristics required a categorical option choice (e.g., gender) whereas others required free-text responses (e.g., height and weight). In line with other studies (Carlson et al., 2008; Zarkadi et al., 2009), we then entered the modal descriptions into the Florida Department of Corrections Inmate Database (http://www.dc.state.fl.us/AppCommon/) to retrieve 40

photographs of men who matched the modal description of each of the four culprits (160 photos in total). This approach fits with the recommendation that foils should match the witness's description of the culprit (Technical Working Group for Eyewitness Evidence, 1999; Wells, 1993).

The photos we selected from the database depicted men facing directly towards the camera. To control for the influence of emotional display, we selected men with neutral facial expressions (Flowe et al., 2014). We used Adobe Photoshop–CS5® to transform the images to grey scale and to remove any background colour or pattern. If the person had a distinctive facial feature, we removed it. To prevent biases attributable to clothing, we also digitally altered each photo so that all foils appeared to be wearing a plain black t-shirt (R. C. L. Lindsay, Wallbridge, & Drennan, 1987). We took similar-looking "mug shots" of the culprits on the day we filmed the mock crimes. We edited these mug shots in the same way as the foil photographs, including adjusting the resolution to match that of the foil photographs.

Next, we edited the four pools of 40 (160 total) images to create foils for the replication, pixelation, and block lineups (see Figure 2.1). For the replication lineups, we digitally added the culprit's distinctive feature to each foil in the pool of 40. To reflect current police practice in several jurisdictions including England, Wales, New Zealand, Canada, and Germany, this distinctive feature was very similar in size, appearance and location—but not identical to—the culprit's distinctive feature. For pixelation lineups, we concealed the culprit's distinctive feature by pixelating it, and pixelating the same region on each of the 40 foils in the corresponding pool. For block lineups, we concealed the culprit's distinctive feature by overlaying a solid black rectangle and we overlaid the same shape, in the same region, on each of the 40 foils in the corresponding pool. For target-present do-nothing lineups, we left the culprit's distinctive feature uncovered and did nothing to the photos of the foils. In target-absent do-nothing lineups, we needed one foil face that had a distinctive feature similar to the culprit's; accordingly, we used one replication foil face to which the culprit's distinctive feature had been added. The other 5 foil photos in each do-nothing target-absent lineup remained undoctored. Note that the do-nothing target-absent lineups mirror the real-world situation in which a witness reports the culprit's distinctive feature to the police, but the police apprehend an innocent person with a similar distinctive feature and place him in the lineup.

To check that we had doctored our foils the way police actually doctor foils, we gathered evidence of ecological validity by consulting with a Detective Inspector from a local police force in the UK who sat on the National Committee for Identification Evidence. We randomly selected 18 foils to whom we had applied the replication, pixelation and block manipulation, and asked her to evaluate them. The officer agreed that the images were concordant with police practice in England and Wales.

To ensure our replication foils did not look doctored, we then asked 5 new subjects to view all four replication foil pools (160 photos) and to identify any images that either did not match the modal description of the culprit, or looked as though they had been digitally altered. These subjects said that all the foils matched the descriptions of the culprits, but identified a total of 14 photos as looking as though they had been digitally altered. We then reedited the distinctive features on these 14 photos until all 5 subjects were satisfied. Next, we asked a new group of 39 subjects to evaluate four target-present replication lineups (one for each culprit), in which the foils were randomly generated. We asked them to identify which photograph had *not* been digitally altered; they were no better than chance at this task (all ps > .20). Taken together, these findings suggest that our replication photos did not look manipulated, and our procedure for generating lineups did not bias subjects towards or against the suspect.

Procedure

Subjects were told that the study was about personality and perception. They were randomly assigned into one of the eight experimental conditions and one of the four crime videos (with the constraint that subject numbers were relatively equal in each condition).

There were three phases in the experiment. In the first phase, subjects watched a video of a crime. They were instructed to pay close attention because they would be asked questions about it later. After the video ended, we asked subjects if they had encountered any technical problems while viewing the video. The second phase, a filler phase, then began. In this phase, subjects worked on three questionnaires and an anagram puzzle for a total of 8 min. The questionnaires were the Autism Spectrum Quotient (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001), the Six-Item Short-Form State scale of the Spielberger State-Trait Anxiety Inventory

(Marteau & Bekker, 1992), and the Ten-Item Personality Inventory (Gosling, Rentfrow, & Swann, 2003). We do not discuss subjects' performance on these scales because they served as a filler task. In the third phase, we asked subjects to indicate their confidence that they would be able to recognise the culprit. Subjects responded on a 100-point Likert-type scale ranging from 1 (*completely uncertain*) to 100 (*completely certain*). Immediately after this task, subjects saw a lineup composed of two rows of three photos. Target-present lineups featured the culprit and five randomly selected foils from the corresponding pool. The position of the culprit was randomly determined for each subject. Replication, pixelation and block target-absent lineups consisted of six randomly selected foils (i.e. there was no designated innocent suspect). In do-nothing target-absent lineups, one foil with the culprit's distinctive feature and five foils without the culprit's distinctive feature were randomly selected (i.e. the innocent suspect was the foil that had the culprit's distinctive feature). The position of the innocent suspect was also randomly determined for each subject. We chose this method of generating lineups to increase the generalizability of our results and to avoid the problems associated with using a small number of culprit and innocent suspect pairs. By randomly generating lineups, we also avoided using lineup fairness and bias measures, which are not always stable (Mansour, Beaudry, Kalmet, Bertrand, & Lindsay, in press).

All subjects were instructed that the culprit "may or may not be present" and then were asked to make a single identification by either clicking on the person they believed to be the culprit, or on an option labelled "Not Present." Next, subjects used a 100-point Likert-type scale (1 = *completely uncertain* to 100 = *completely certain*) to rate their confidence in their decision. Finally, subjects answered a question that enabled us to check that they were paying attention ("What happened in the video that you watched?"), and they also answered a number of demographic questions.

Results

Recall that our primary aim was to determine the extent to which unfair lineups affect witnesses' (a) willingness to identify the suspect, and (b) ability to distinguish between real culprits and innocent suspects. We addressed these questions by using ROC analysis, and gathered further information by examining the distribution of subjects' identification responses and subjects' ability to judge the accuracy of their identification decisions.

ROC analysis

To be clear, our ROC analysis measures people's ability to discriminate between guilty and innocent suspects, setting aside choices of known-to-be innocent foils. This is different to an absolute notion of memory discriminability—which would be the ability to discriminate between guilty suspects and anyone else in the lineup (i.e., innocent suspects and foils; see Wixted & Mickes, 2015 for a discussion). From a practical standpoint, discriminating between guilty and innocent suspects is arguably the key discriminability to measure because false identifications of foils do not result in any legal action against the foil that is selected. Nevertheless, we direct interested readers to our signal detection modelling reported in Appendix A, because the modelling accounts for foil choices. That is, the modelling also estimates people's ability to discriminate (a) guilty suspects from foils and (b) innocent suspects from foils in unfair lineups.

To construct our ROC curves, we collapsed the data across the four crime videos. We rounded subjects' confidence ratings (made on a 100-point Likert scale) to the nearest 10 so that each curve would have 11 operating points of decreasing confidence (i.e. 100, 90, 80, and so forth). We then calculated the correct identification rates (hit rates; HRs) and the false identification rates (false identification rates; FARs) over the decreasing confidence levels. The correct identification rate (HR) was the number of guilty suspect IDs ÷ number of target-present lineups. The false identification rate (FAR) was the number of innocent suspect IDs ÷ number of target-absent lineups (see Chapter 1, or Mickes et al., 2012, for more comprehensive tutorials on ROC analysis).

We calculated innocent suspect identifications differently for the unfair and fair lineups. In the unfair (do-nothing) lineups, subjects made innocent suspect identifications when they identified the single lineup member with the distinctive feature. In the fair (replication, pixelation and block) lineups, recall that there was no designated innocent suspect—thus we estimated the number of innocent suspect identifications in these conditions using a common approach. We divided the number of false identifications made in target-absent lineups by the total number of people in the lineup—here, six (Brewer & Wells, 2006; Mickes, 2015). This procedure works on the assumption that the lineup member that best matches the subject's memory of the culprit is the innocent suspect (Palmer et al., 2013). One

particular benefit of estimating false identifications in this way is that it leads to a more conservative measure of false identifications. Because the innocent suspect may not always be the most similar in appearance to the actual culprit, this method of estimation can only overestimate, not underestimate, the number of false identifications in target-absent lineups. Thus, using this estimation method in replication, pixelation and block lineups provided a conservative test of how well these (fair) techniques enhance witness identification performance compared to the (unfair) do-nothing lineups.

To calculate *p*AUC, we used the statistical package pROC (Version 1.8; Robin et al., 2011) with RStudio (Version 0.98.1103; RStudio Team, 2015) and the R software environment (Version 3.2.0; R Development Core Team, 2015). pROC also calculates a measure of effect size, *D*, using the formula: $D = (AUC1 - AUC2)/s$. In this formula, *s* is the standard error of the difference between the two AUCs and is estimated using bootstrapping.

Figure 2.2 shows the ROC curves for the fair and unfair lineups. When calculating *p*AUC statistics, we set the specificity to .91—which corresponded to the FAR range covered by the least extensive curve (block; FAR range: 0 to .09)—for two main reasons. First, by setting the FAR range from 0 to .09, we prevented the pROC program from having to extrapolate the three fair lineup curves over a vast range (from a FAR of .09 to .40). The pROC program uses a crude method of extrapolation, so doing so over large distances can reduce statistical accuracy. Second, the lower FAR range (0 to .09) may have greater practical relevance, because the legal system (a) is interested in knowing which conditions increase witnesses' ability to distinguish between innocent and guilty suspects when the FAR is low, and (b) may take these high-confidence identifications more seriously than low-confidence identifications (see Gronlund et al., 2012). We are confident that limiting the *p*AUC analysis to a small subset of the do-nothing curve did not affect our findings. When we fit a theoretical model to our data we found the same pattern of results (see Appendix A). This modelling technique uses the largest FAR range that a target-absent lineup can support.

To what extent did our lineup types affect witnesses' performance? More specifically, did the unfair lineups increase witnesses' willingness to choose the suspect—or did those lineups impair witnesses' ability to distinguish between guilty

and innocent suspects? As Figure 2.2 shows, compared to the replication, pixelation and block (i.e., fair) lineup techniques, doing nothing increased subjects' willingness to identify the suspect and also markedly impaired subjects' ability to discriminate between real culprits and innocent suspects. Focusing on the ROC curves in Figure 2.2, we can see that the do-nothing ROC points have shifted more to the right than any of the fair lineup ROC points. This shift right shows there was an increase in both correct and false identifications. That is, subjects' willingness to identify the suspect increased in the do-nothing lineups, as compared to replication, pixelation and block lineups.

A more striking finding though, is that do-nothing lineups made it more difficult for subjects to distinguish between innocent and guilty suspects. The *p*AUC for do-nothing lineups (*p*AUC = 0.008, 95% CI: 0.006, 0.010) was significantly smaller than the *p*AUC for replication (*p*AUC = 0.016, 95% CI: 0.013, 0.019, $D = 4.11$, $p < .001$), pixelation (*p*AUC = 0.015, 95% CI: 0.012, 0.018, $D = 4.17$, $p < .001$) and block (*p*AUC = 0.016, 95% CI: 0.013, 0.019, $D = 4.35$, $p < .001$) lineups. Finally, the three fair lineups led to similar levels of identification performance—the *p*AUCs did not differ significantly between replication and pixelation ($D = 0.32$, $p > .250$), replication and block ($D = 0.08$, $p > .250$), or pixelation and block ($D = 0.24$, $p > .250$) lineups. We also fit a signal detection process model to our data to further confirm these findings (see Appendix A). Importantly, the model-fitting exercise and our *p*AUC analysis led to the same results. Taken together, these findings fit with the additional prediction of the diagnostic-feature-detection model—that doing nothing to stop distinctive suspects from standing out does not just increase witnesses' willingness to choose the suspect, it also markedly impairs their ability to sort guilty and innocent suspects into their appropriate categories.

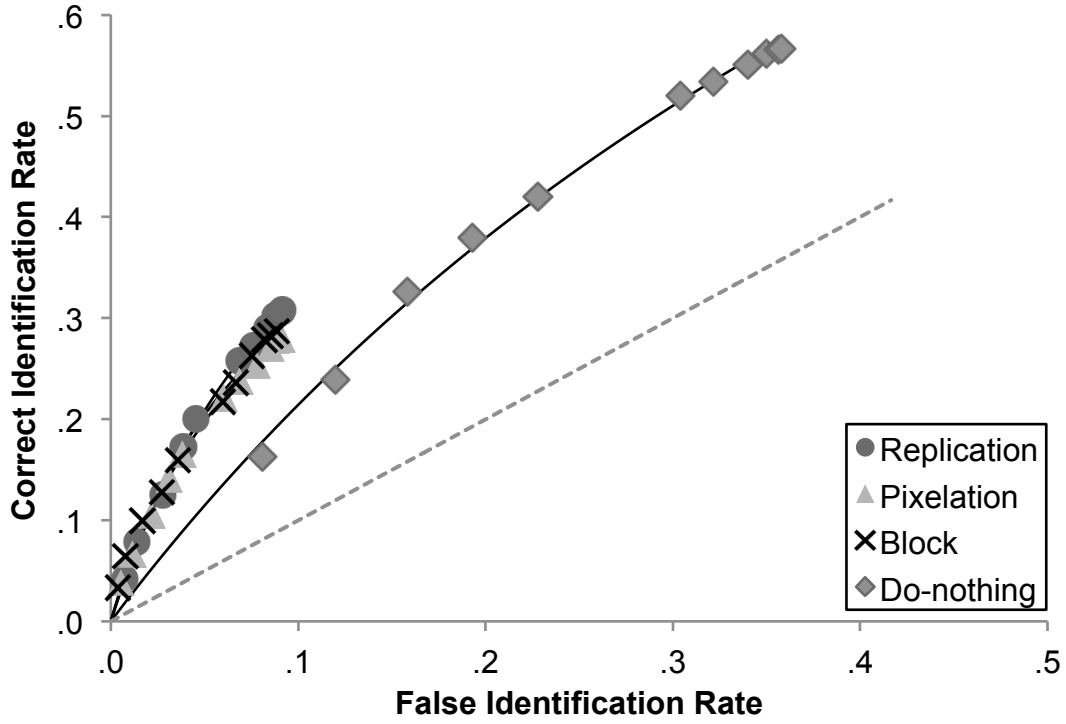

Figure 2.2. Receiver operating characteristic (ROC) curves for the fair (replication, pixelation, block) and unfair (do-nothing) lineups. The dashed line represents chance-level performance.

Identification responses

To further understand the effect of unfair lineups on subjects' identification performance, we calculated the proportion of suspect identifications, foil identifications and lineup rejections (i.e., "Not Present" responses) for each lineup type. Table 2.2 shows the frequencies and percentages of identification responses for each lineup type. There is an interesting point to note about these data. We know from the ROC analysis that unfair lineups led to more (guilty and innocent) suspect identifications than did fair lineups. The data in Table 2.2 indicate that this overall increase in suspect identifications was accompanied by a decrease in both foil identifications and lineup rejections in target-present lineups, but just a decrease in foil identifications in target-absent lineups.

Table 2.2
Frequencies and Percentages of Identification Responses in the Replication, Pixelation, Block, and Do-nothing Lineups

Identification responses	Replication		Pixelation		Block		Do-nothing	
	n	%	*n*	%	*n*	%	*n*	%
Target present								
Guilty suspect	347.00	30.84	320.00	27.95	323.00	28.66	629.00	56.67
Foil	382.00	33.96	411.00	35.90	390.00	34.61	206.00	18.56
Incorrect rejection	396.00	35.20	414.00	36.16	414.00	36.73	275.00	24.77
Target absent								
Innocent suspect	104.50	9.17	102.33	9.10	100.50	8.84	364.00	35.79
Foil	522.50	45.83	511.67	45.52	502.50	44.20	219.00	21.53
Correct rejection	513.00	45.00	510.00	45.37	534.00	46.97	434.00	42.67

Target-present lineups. A 4 (lineup type: replication, pixelation, block, do-nothing) × 3 (identification response: guilty suspect, foil, incorrect rejection) chi-square analysis showed that lineup type influenced ID responses, χ^2 (6, N = 4,507) = 282.70, $p < .001$, Cramer's $V = .18$. Specifically, fair lineups led to fewer guilty suspect IDs (replication: $z = –2.84$, $p < .01$; pixelation: $z = –4.50$, $p < .001$; block: $z = –4.07$, $p < .001$) but more foil IDs (replication: $z = 1.90$, $p > .05$; pixelation: $z = 3.09$, $p < .01$; block: $z = 2.29$, $p < .05$) and more lineup rejections (replication: $z = 1.13$, $p > .05$; pixelation: $z = 1.70$, $p > .05$; block: $z = 2.02$, $p < .05$) than expected. Conversely, unfair lineups led to more guilty suspect IDs ($z = 11.53$, $p < .001$), but fewer foil IDs ($z = –7.36$, $p < .001$) and fewer lineup rejections ($z = –4.90$, $p < .001$) than expected. In short, when the suspect was left to stand out in target-present lineups, there was an increase in guilty suspect identifications along with a reduction in both foil identifications and incorrect rejections.

Target-absent lineups. Recall that in replication, pixelation and block target-absent lineups there was no designated innocent suspect. We therefore estimated the number of innocent suspect identifications by dividing the total number of false identifications by six (the number of faces in the lineup). Similarly, we estimated the number of foil identifications by dividing the total number of false identifications by six (the number of faces in the lineup), and then multiplying by five (the number of faces that were not the innocent suspect). A 4 (lineup type: replication, pixelation, block, do-nothing) × 3 (identification response: innocent suspect, foil, correct

rejection) chi-square analysis using these estimates showed that lineup technique influenced ID responses, χ^2 (6, N = 4,418) = 481.70, p < .001, Cramer's V = .23. Fair lineups led to fewer innocent suspect IDs (replication: z = –5.22, p < .001; pixelation: z = –5.24, p < .001; block: z = –5.50, p < .001), but more foil IDs (replication: z = 3.26, p < .001; pixelation: z = 3.08, p < .001; block: z = 2.38, p < .001) than expected. Conversely, unfair lineups led to more innocent suspect IDs (z = 16.85, p < .001), but fewer foil IDs (z = –9.21, p < .001) than expected. The proportion of correct lineup rejections in all four lineup types was similar (replication: z = –0.03, p >.05; pixelation: z = 0.15, p > .05; block: z = 0.954, p > .05; do-nothing: z = –1.14, p >.05). This analysis indicates that when the suspect was left to stand out in target-absent lineups, subjects shifted their identifications from the other lineup members, onto the innocent suspect.

Confidence and accuracy

Recall that the diagnostic-feature-detection model suggests that unfair lineups impair a witness's ability to distinguish between innocent and guilty suspects because it is not obvious to the witness that the suspect's distinctive feature is unhelpful. If witnesses fail to realise that the distinctive feature is unhelpful, they may not lower their confidence judgement to compensate for their poorer performance. If this account is correct, then subjects who viewed the unfair do-nothing lineups should be less accurate, at every level of confidence, than subjects who viewed the fair replication, pixelation and block lineups.

To test this prediction, we plotted suspect identification accuracy (correct IDs of guilty suspects in target-present lineups ÷ [correct IDs of guilty suspects in target-present lineups + false IDs of innocent suspects in target-absent lineups]) separately for each level of confidence (100, 90, 80, and so forth, as per Mickes, 2015). This method of calculating suspect-identification accuracy reflects the probability of guilt, given that the suspect was identified (i.e., the posterior probability of guilt). We estimated the number of innocent suspect identifications in the replication, block and pixelation lineups in the same way we did for the ROC analysis. To provide more stable estimates, confidence level was binned into five categories (0–20, 30–40, 50–60, 70–80, 90–100, see Brewer & Wells, 2006). The frequencies of identification responses in each confidence bin are presented in Appendix B.

Figure 2.3 shows the confidence-accuracy curves for each lineup type. Non-overlapping error bars denote reliable differences between the lineup techniques (e.g. Sauer et al., 2010). As predicted, subjects who viewed the unfair, do-nothing lineups showed lower levels of accuracy at every level of confidence than subjects who viewed the fair lineups. Put another way, an identification made at any level of confidence from an unfair lineup was less trustworthy than an identification made with the same level of confidence from a fair lineup. These data align with the diagnostic-feature-detection model, which suggests that when nothing was done to stop the distinctive suspect from standing out, subjects may have been unaware that their memory accuracy was worse and therefore failed to adjust their confidence accordingly.

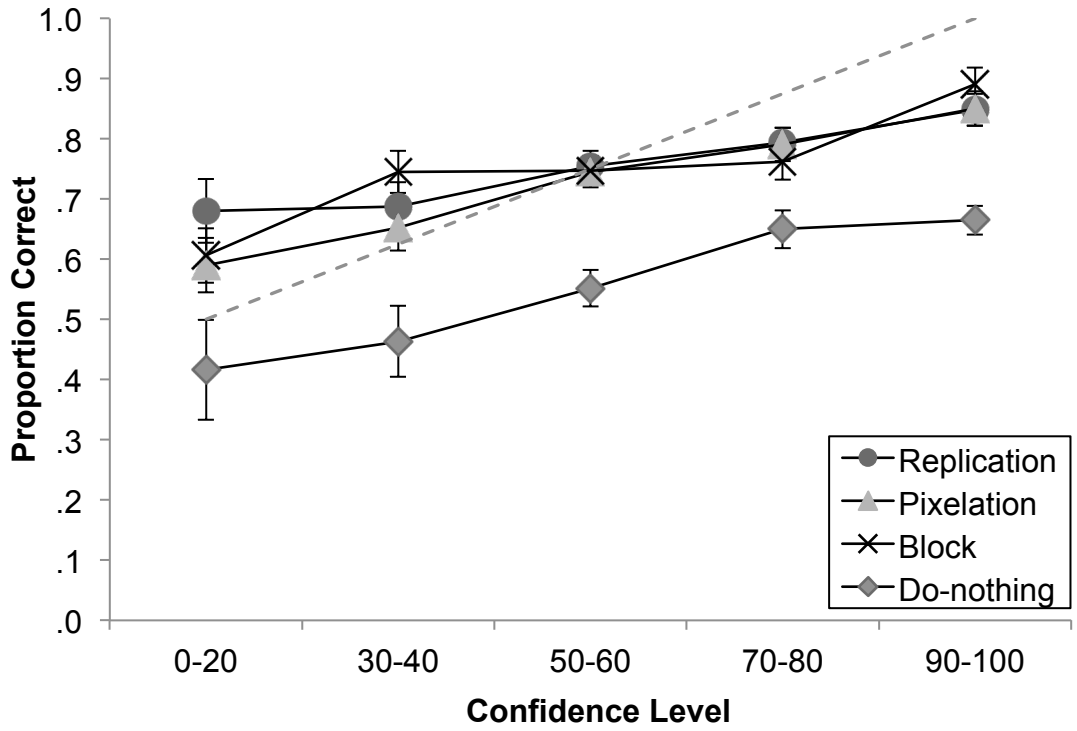

Figure 2.3. Confidence-accuracy curves for suspect identifications in the fair (replication, pixelation, block) and unfair (do-nothing) lineups. Error bars indicate ±1 SE. The dashed line represents chance accuracy at the lowest confidence bin (i.e., 0–20) and perfect accuracy at the highest confidence bin (i.e., 90–100).

Discussion

We asked why unfair lineups promote mistaken identifications. Our findings suggest that unfair lineups—in comparison to fair lineups—make people more likely

to identify the suspect, but, worse still, unfair lineups impair people's ability to distinguish between guilty and innocent suspects and distort people's ability to judge the trustworthiness of their identification decision.

It is arguably unsurprising that our unfair lineups, in which a suspect was left to stand out, increased subjects' willingness to identify that suspect. Many eyewitness identification studies have demonstrated this already (Clark, 2012; Doob & Kirshenbaum, 1973; Fitzgerald et al., 2013; Wells, Leippe, & Ostrom, 1979; Wells et al., 1993). The fascinating finding is that unfair lineups also dramatically hindered subjects' ability to sort innocent and guilty suspects into their appropriate categories. This mechanism has not been discussed until now, yet it is important. Procedures that simply make witnesses less willing to choose the suspect decrease innocent suspect identifications but also come at a cost: they stifle culprit identifications (Clark, 2012). Procedures that enhance a witness's ability to distinguish between innocent and guilty suspects minimise innocent suspect identifications *and* maximise culprit identifications, regardless of the witness's willingness to choose. Arguably then, this is the critical mechanism to investigate (Gronlund et al., 2014; National Research Council, 2014).

So, why might unfair lineups harm people's ability to distinguish between the real culprit and an innocent suspect? One explanation is that witnesses fail to appreciate that the suspect's distinctive feature is not useful in an unfair lineup and so rely heavily on it to make their identification. By contrast, when lineups are fair and the suspect does not stand out, witnesses can appropriately discount the distinctive feature and give more weight to other, more informative, cues (Wixted & Mickes, 2014). Support for this theoretical account comes from the finding that, in the unfair lineups, subjects failed to compensate by setting a more conservative confidence criterion when making an identification. This fits with a mechanism in which subjects do not realise that their accuracy is impaired.

Importantly, a growing body of research suggests that subjects acting as witnesses in studies are generally good at judging the likely accuracy of their memories even when their accuracy is impaired (e.g. Brewer & Wells, 2006; Mickes, 2015, Experiment 1; Palmer et al., 2013; Sauer et al., 2010). Palmer et al., for instance, showed that divided attention significantly impaired people's memory ability, yet, when the authors plotted accuracy at each level of confidence it didn't

matter if subjects had full or divided attention at encoding—subjects' accuracy at each level of confidence was generally the same (Experiment 2, Figures 3 and 4). Palmer and colleagues concluded that their experimental manipulations did not undermine the usefulness of confidence as an indicator of accuracy. This study, and many others, shows that people typically recognise when their memories are poor and adjust their confidence appropriately (Mickes, 2015, Experiment 1; Palmer et al. 2013, Experiment 1; Sauer et al., 2010). There are, however, some instances in which confidence is uninformative of accuracy (e.g., Chandler, 1994; Mickes, 2015, Experiment 2). Indeed, our findings show that unfair lineups can systematically distort confidence.

One consequence of identifications from unfair lineups being less accurate at every level of confidence, is that subjects in the do-nothing condition made high-confidence suspect identifications (certainty of 90–100) when accuracy was moderate (.60). This finding has serious implications for criminal justice because legal decision makers are strongly influenced by highly confident witnesses (Brewer & Burke, 2002; Wells, Lindsay, & Ferguson, 1979). Although subjects in the fair lineup conditions (i.e., replication, pixelation or block) were under-confident at the lower end of the confidence scale, the critical point is that their identifications were consistently and substantially more trustworthy than the identifications made by subjects in the unfair lineup condition. Moreover, subjects who viewed the fair lineups identified the suspect with high confidence (certainty of 90–100) when they were very likely to be accurate (> .80). Therefore, highly-confident suspect identifications made from replication, pixelation and block lineups are likely to be very informative for triers of fact.

At first glance, our results appear to conflict with two face-recognition studies that suggest replicating distinctive features is better than removing them (Badham et al., 2013; Zarkadi et al., 2009). Zarkadi and colleagues, for example, found that replication increased correct identifications by approximately 20% in target-present lineups, while we found replication and concealment techniques were equally effective. There is, however, a crucial methodological difference to consider. The previous research compared replication lineups with removal lineups in which the target's distinctive feature was simply removed. Subjects made more incorrect rejections in target-present removal lineups possibly because the person they

believed to be the culprit was now missing a prominent distinctive feature that they remembered (Wixted & Mickes, 2014). Subjects in our study were unlikely to use this strategy because we tested pixelation and block lineups, both of which indicate that there could be a distinctive feature underneath the concealed area. Therefore, unlike the previous research, we did not observe a relatively high number of incorrect rejections in pixelation and block lineups compared to replication lineups. Instead, we observed similar performance in all three fair conditions.

On a practical level, our research suggests that law enforcement officers should take steps to prevent distinctive suspects from standing out. If unfair lineups just increased witnesses' willingness to choose the suspect (and did not affect their ability to distinguish between innocent and guilty suspects), then officers could remedy this by inducing more conservative responding. For instance, urging witnesses to be cautious ("Be certain before making a decision") should increase the amount of memory information that witnesses demand before choosing and result in fewer positive, and therefore fewer suspect, identifications (Clark, 2005). Our data, however, suggest that law enforcement officers need to apply fair lineup techniques to improve identification accuracy, and that replication, pixelation, or block techniques are equally effective.

In sum, our data fit the predictions of a new model, the diagnostic-feature-detection model. Testing theoretical models is important, because, once refined, theories can be used to develop procedures that further enhance eyewitness accuracy. More specifically, our findings shed light on the processes underlying the harmful effects of unfair lineups and suggest that when suspects are unduly distinctive, witnesses are not just more willing to choose the suspect, they also struggle to distinguish between guilty and innocent suspects. Perhaps if Leonard Callace had been placed in a fair lineup, alongside foils who also had full beards or whose chins had been concealed, he would not have spent 6 years in prison for a crime he did not commit.

Chapter 3 :
Identification Performance and Age

"…the memory of a witness may fade, particularly when, as in this case, the witness is elderly."

Niblett v. Commonwealth (1976)

Overview

To construct fair lineups for suspects with distinctive features (e.g., scars, birthmarks), police officers can use one of three techniques—replication, pixelation or block—to prevent suspects from standing out. In Chapter 2, we found that all three fair lineups techniques for distinctive suspects were equally effective at enhancing people's ability to discriminate between innocent and guilty suspects compared to unfair (do-nothing) lineups in which the suspect was left to stand out. According to the diagnostic-feature-detection account, all three fair lineups elicit similar identification performance because all three lineups encourage witnesses to discount the distinctive feature, and, instead, encourage reliance on other facial features that are diagnostic of guilt (Wixted & Mickes, 2014).

But are all three fair lineup techniques equally effective in witnesses of all ages? Although the diagnostic-feature-detection model predicts that all three fair lineups should elicit similar identification performance, this prediction may not hold for older witnesses. When making memory decisions, older adults are more likely to rely on a feeling of familiarity, rather than recollecting specific details (e.g., Healy, Light, & Chung, 2005; Searcy, Bartlett, & Memon, 1999). In replication lineups, replicating distinctive features across foils might make those foils seem more familiar, because the foils all share a similar distinctive feature with the culprit. By contrast, in pixelation and block lineups, concealing the relevant area on the foils does not make the foils more familiar, because the foils simply have an area that has been pixelated or blocked out. If older adults are overly reliant on familiarity when making their identification decision, and all of the faces feel more familiar in replication lineups, then older adults may make more incorrect identifications of foils in replication lineups compared to pixelation or block lineups.

We investigated the efficacy of the fair replication, pixelation, and block lineup techniques in young (18–30 years, n = 890), middle-aged (31–59 years, n = 890) and older (60–95 years, n = 890) adults by examining subjects' identification responses, conducting ROC analysis and fitting a signal detection process model to our data (Wixted & Mickes, 2014, see Appendix C for these analyses). We replicated our findings from Chapter 2—within each age group, all three fair lineups led to similar identification performance. Practically, these findings illustrate that there are multiple ways to construct fair lineups for distinctive suspects, in young, middle-aged and older adults.

The large dataset, coupled with the theoretical and statistical techniques outlined in Chapter 1, also provided the opportunity to examine general changes in identification performance across the three age groups. Our analyses resulted in a number of novel findings that have important implications for interpreting identifications made by middle-aged and older witnesses. Therefore, we used the data to publish a more general paper that focused on how identification performance changes—in fair and unfair lineups—with healthy ageing. That paper (currently under review at *Psychology and Aging*) is presented in this chapter. Information that has been covered in previous chapters has been edited to avoid repetitiveness.

Introduction

Imagine that you are a police officer investigating a crime. You have only one witness, a 69-year-old, whose ability to recognise the culprit is critical for your case. How might your witness's ability to make an accurate identification from a lineup be different to that of a young or middle-aged adult? Now imagine that you are a judge deliberating the verdict. Can you trust the identification made by this older witness to the same extent that you might trust an identification made by a younger witness? In nearly every country, the proportion of people aged 60 and over is growing faster than any other age group (World Health Organization, 2015), and middle-aged and older adults are frequently witnesses or victims of crime (e.g., Acierno et al., 2010; Willoughby, 2015). Yet, knowledge of how eyewitness identification performance changes with age is limited (Fitzgerald & Price, 2015). In this study, we aimed to learn more about eyewitness identification behaviour in middle-aged and older adults by examining their ability to identify culprits and gauge the accuracy of their identification decisions.

Many eyewitness identification studies have shown that older adults make more mistakes in lineup tasks than do young adults. Older adults are more likely than young adults, for instance, to make an incorrect identification when the real culprit is not in the lineup (see Bartlett & Memon, 2007 for a review). Early studies also found that older adults are more likely to select a person from a lineup than are their young counterparts (see Sporer & Martschuk, 2014 for a review). As a result many researchers have, explicitly or implicitly, suggested that the age-related decline in identification accuracy occurs because older adults are too willing to make an identification decision (e.g., Sporer & Martschuk, 2014; Wilcock, Bull, & Vrij, 2005). However, attempts to reduce older adults' false identification rates—by reducing proclivity to choose—have not been effective in eradicating the age-related deficit in performance (Memon & Gabbert, 2003; Rose, Bull, & Vrij, 2005; Wilcock et al., 2005). It seems that an increased willingness to choose with age is not the whole story.

Indeed, there are good reasons to expect that ageing is associated with a genuine decline in recognition accuracy—also known as *discriminability*—and not just an increased willingness to choose. Healthy ageing is associated with a number of changes in memory function, but one prominent theory suggests that people become increasingly reliant on familiarity with age and this tendency promotes memory errors (Healy et al., 2005; Searcy et al., 1999). According to dual-process accounts of memory, recognition is based on two processes: recollection and familiarity (see Mandler, 1980 and Yonelinas, 2002 for reviews). Recollection involves retrieving specific contextual information about the original stimulus, such as source, time, place, thoughts and feelings, whereas familiarity is a sense that the stimulus has previously been encountered without retrieving any contextual details. Evidence from several different paradigms including old/new word recognition studies (Dywan & Jacoby, 1990; Jacoby, 1999; Jennings & Jacoby, 1997), face recognition studies (Bartlett & Fulton, 1991; Bartlett, Strater, & Fulton, 1991; Edmonds, Glisky, Bartlett, & Rapcsak, 2012), and lineup tasks (Searcy et al., 1999, Searcy, Bartlett, & Memon, 2000; Searcy, Bartlett, Memon, & Swanson, 2001), suggest that older adults have deficits in recollecting diagnostic source specific information and, as a result, are more reliant on less diagnostic familiarity processes than are their younger counterparts.

What does this mean for older adults' ability to discriminate between who is innocent and who is guilty in a lineup? Faces in a lineup are highly homogenous (Diamond & Carey, 1986), so even faces that have never been seen before could evoke a feeling of familiarity (Bartlett, Hurry, & Thorley, 1984; Young, Hay, McWeeny, Flude, & Ellis, 1985). Because older adults are poorer at recollecting diagnostic details associated with a previously seen face, they may rely on familiarity to a greater extent than young adults, thereby making it harder for them to tell if a person in the lineup is innocent or guilty.

Indeed, face recognition studies show that discriminability declines with age (e.g., Fulton & Bartlett, 1991; Lamont, Stewart-Williams, & Podd, 2005). Three meta-analyses of lineup research have shown that, compared to young adults, older adults make more false identifications when the culprit is not in the lineup, but also fewer correct identifications when the culprit is in the lineup (Bartlett, 2014; Fitzgerald & Price, 2015; Sporer & Martschuk, 2014). Only three studies, however, have directly measured young and older adults' ability to discriminate between innocent and guilty suspects as well as their willingness to identify the suspect. One study calculated overall choosing rate and signal detection estimates of discrimination (d') and response bias (c) for 21 published lineup studies. The authors concluded that while older adults do choose from lineups at a higher rate than young adults, it was an impaired ability to discriminate between innocent and guilty suspects that hindered older adults' performance (Fitzgerald & Price, 2015; see also Wylie, Bergt, Haby, Brank, & Bornstein, 2015). By contrast, Key et al. (2015) measured people's ability to discriminate between innocent and guilty suspects in fair lineups (where the foils matched the appearance of the suspect) and unfair lineups (where the suspect stood out because the foils did not match the appearance of the suspect) using ROC analysis. Surprisingly, Key et al. found no difference between their young and older samples on either lineup type.

If people's ability to discriminate between innocent and guilty suspects declines with age, should the Criminal Justice System disregard identifications made by older, or even middle-aged, adults? Somewhat surprisingly, merely knowing that older adults have lower discriminability does not provide us with the information needed to answer that question. To answer that question, we need to consider whether older adults can assess the likely accuracy of their memories and assign

appropriate confidence judgements (Mickes, 2015). That is, do older adults express high confidence in their decision when their answer is correct, and lower confidence when their answer is incorrect, and do they do so to the same degree as younger people? If they do, then a high-confidence ID from an older adult would be as trustworthy as a high-confidence ID from a younger adult even though older adults exhibit reduced discriminability.

Gauging the accuracy of identifications

Currently, the eyewitness research on confidence judgements in older adults is mixed. Some lineup studies have found that accuracy and confidence are better correlated in young people than in older people (Adams-Price, 1992; Memon, Hope, Bartlett, & Bull, 2002; Wylie et al., 2015), and a recent review of this lineup literature concluded that confidence should not be used as a proxy for accuracy in older adults (Erickson, Lampinen, & Moore, 2015). Also, older adults often make high-confidence errors (Dodson, Bawa, & Krueger, 2007; Dodson, Bawa, & Slotnick, 2007; Dodson & Krueger, 2006), and older adults who rate their memory self-efficacy as higher are more likely to make false identifications (Searcy et al., 2000, 2001). These studies may indicate that older adults tend to be over-confident in the validity of weaker memory signals. That is, older adults may not adjust their confidence judgements appropriately to reflect their lower likelihood of accuracy.

However, it may be premature to conclude that older adults are unable to assign appropriate confidence judgements. Many of the lineup studies (e.g., Adams-Price, 1992; Memon et al., 2002; Wylie et al., 2015) have calculated the correlation coefficient, but we now know that a low correlation coefficient does not necessarily indicate a poor relationship between confidence and accuracy (Juslin et al., 1996). A more suitable statistical technique for testing whether people can assess the likely accuracy of their memories is to plot their *average* accuracy at different levels of confidence—that is, plot confidence-accuracy curves. Only this technique tells us the likely accuracy of an identification made with a particular level of confidence (Brewer et al., 2002; Brewer & Wells, 2006; Juslin et el., 1996; Mickes, 2015; see also Chapter 1).

To our knowledge, Key et al. (2015) is the only study to have plotted confidence-accuracy curves for young and older adults in an eyewitness identification paradigm. When older adults made suspect identifications with the

highest level of confidence, they were as likely to be correct as were young adults. This finding should be interpreted with caution, though, because the young and older groups were also equivalent in discriminability. What this study does not tell us, then, is whether older adults can assess the accuracy of their memories to the same extent as young adults, even when their memory ability is worse. Nevertheless, many other eyewitness studies have found that older adults tend to assign lower confidence ratings to their identification decisions on average than young adults, which may suggest that older adults are aware that they are less accurate (Goodsell, Neuschatz, & Gronlund, 2009; Memon et al., 2002; Neuschatz et al., 2005; Searcy et al., 2001; Wylie et al., 2015; but see Havard & Memon, 2009; Searcy et al., 1999). If middle-aged and older adults are able to gauge the likely accuracy of their memories, then they should be as accurate as young adults at each level of confidence, despite any decline in memory ability that occurs with age.

Current study

We aimed to answer two main questions: [1] Is the age-related decline in accurate identification decisions due to an increased willingness to make an identification, a decline in discriminability, or both? [2] Are middle-aged and older adults able to gauge the likely accuracy of their suspect identification decisions to the same extent as young adults? To answer these questions, young (18–30 years), middle-aged (31–59 years) and older (60–95 years) adults watched a video of a mock crime and attempted to identify the culprit from a lineup. The lineup was either fair, in which all of the lineup members matched the appearance of the suspect, or unfair, in which the suspect stood out. Subjects also provided confidence ratings for their identification decisions. We conducted ROC analysis and plotted confidence-accuracy curves. To further understand the mechanisms underlying the trends in identification responses, we also fit a signal detection process model of eyewitness identification behaviour to our data.

Method

Design

ROC analysis requires a large data set to ensure stable ROC functions. To aid our data collection we combined a subset of the data from Chapter 2 with newly collected data. Data collection for both studies occurred within a nine-month period.

We tested our subjects using the same stimuli and procedure as in Chapter 2. Therefore, we used a 3 (age: young, middle-aged, older) × 4 (lineup type: replication, pixelation, block, do-nothing) × 2 (target: present, absent) between-subjects design. Because we found no difference between performance on the replication, pixelation and block lineups in Chapter 2, we planned to collapse the data over the three fair lineup techniques.

Subjects

Older adults. Our data collection stopping rule was to recruit as many older subjects as possible before the end of spring term, with a minimum of 60 subjects in each of the eight lineup conditions. To this end, we collected data from 1,285 subjects aged over 60 by contacting University of the Third Age groups from around the UK. Subjects were not paid for their time, but were offered the chance to learn about the research process and the results. Subjects completed the study online and followed the exact same procedure as Chapter 2. We excluded subjects who failed to report their age (n = 4), experienced technical difficulties (n = 38), stated they had seen the video before (n = 9), or incorrectly answered an attention check question on the content of the video (n = 10). Chapter 2 included 8,925 subjects aged between 16 and 91. Of these, 346 subjects were aged over 60. We added these to our cleaned older sample (n = 1,224) to make a total of 1,570 older adults.

Young and middle-aged adults. We sampled 1,570 people aged 18–30 and 1,570 people aged 31–59 from the data collected in Chapter 2. We matched the young and middle-aged samples with our older sample on self-reported sex and ethnicity to ensure that there was a similar number of male and female subjects of each ethnicity in the eight lineup conditions and four mock crime videos (*carjacking, graffiti, mugging, theft*) in all three age groups. If there was no young or middle-aged subject to match an older subject on both ethnicity and sex in a specific cell of the design, we selected an individual of the same ethnicity, disregarding sex (0.64% of young and 2.29% of middle-aged subjects were selected in this way). If there was no young or middle-aged subject to match an older subject on ethnicity, after disregarding sex, then we selected an individual of the same sex who was either: (a) of any non-white ethnicity if the older subject had reported a non-white ethnicity (0.06% of young and 0.38% of middle-aged subjects were selected in this way), or (b) of any ethnicity (white or non-white) if the older subject had chosen to answer

"prefer not to say" or "other" (0.51% of young and 1.08% of middle-aged subjects were selected in this way). Except for this matching process, the young and middle-aged samples were randomly selected.

Although we initially planned to analyse the data from all four videos, in the end, we only analysed the data from the *graffiti* and *mugging* videos because identification performance was very low for the other two videos even for young subjects tested using fair lineups (*carjacking* $d' = 0.74$; *theft* $d' = 0.43$). For older subjects, performance was on the floor (*carjacking* $d' = -0.04$; *theft* $d' = -0.24$). Identification performance on the fair lineups was much better in the *graffiti* (young $d' = 1.21$, middle-aged $d' = 0.96$, older $d' = 0.65$) and *mugging* (young $d' = 1.08$, middle-aged $d' = 1.16$, older $d' = 0.70$) videos. Limiting the analysis to those two videos resulted in a final sample size of 890 older adults (163 from Chapter 2, and 727 new recruits) and 890 middle-aged adults and 890 young adults from Chapter 2. Table 3.1 shows a demographic breakdown of the final sample. Each cell of the design contained between 89 and 117 subjects.

Table 3.1
Demographic Information For Young, Middle-aged, and Older Samples

	Young	Middle-aged	Older
Sex			
Male	311	292	307
Female	579	598	583
Age (years)			
M	22.48	42.49	68.82
SD	3.70	8.27	6.41
Range	18–30	31–59	60–95
Race or ethnicity			
White or European	856	861	853
Latin or Hispanic	1	1	0
Black, African, or Caribbean	9	9	8
South Asian	5	8	5
East Asian	0	1	0
Middle Eastern	1	0	1
Mixed	5	5	6
Other	3	3	3
Prefer not to say	10	2	14

Materials

Videos

We used the four 30-s, non-violent mock crime videos from Chapter 2. However, as stated previously, identification performance was very low on the carjacking and theft videos, so we only included subjects who had watched the graffiti and mugging videos in our analyses.

Lineups

We used the same lineup materials and construction strategy that we used in Chapter 2.

Procedure

We used the same eyewitness memory procedure that we used in Chapter 2.

Results

We examined subjects' identification responses, conducted ROC analysis and fit a signal detection model to our data (Wixted & Mickes, 2014). We also plotted confidence-accuracy curves.

Preliminary analyses

Older adults. We recruited the majority of our older adults from an organisation that promotes lifelong learning. To check that we did not have an unusually able older adult sample, we examined whether our older adults showed the expected speed deficits in performance that accompany normal ageing. Recall that our filler task consisted of three questionnaires followed by an anagram puzzle.[1] The proportions of young, middle-aged and older adults who were still working on the questionnaire items at the end of the 8-min filler task were .10, .13, and .48, respectively. A 3 (age: young, middle-aged, older) × 2 (complete: yes, no) two-way chi-square analysis indicated that completion of the questionnaire items was dependent on age, χ^2 (2, N = 2,645) = 432.68, p < .001, Cramer's V = .40. Specifically, older adults were over 6 times more likely than the middle-aged adults, χ^2 (1, N = 1,755) = 256.16, p < .001, OR = 6.17, 95% CI [4.85, 7.89], and over 8

[1] Due to a technical error, we had missing filler task data from 25 older adults.

times more likely than the young adults, χ^2 (1, N = 1,755) = 307.05, p < .001, OR = 8.10, 95% CI [6.26, 10.57], to still be working on the questionnaire items at the end of the 8-min filler task. Young and middle-aged adults were equally likely to be working on the questionnaire items, χ^2 (1, N = 1,780) = 3.39, p = .07, OR = 1.31, 95% CI [0.97, 1.78].

Recall also that in the experimental task, subjects were asked to make an identification decision from a lineup and then rate their confidence in their decision. A one-way ANOVA showed that the length of time (s) to make an identification decision from the lineup was dependent on age group, F (2, 2667) = 43.69, p < .001. Older adults, M = 17.61, SD = 11.38, were slower than both middle-aged, M = 13.60, SD = 9.91, t (1745.3) = 7.94, p < .001, r = .19, and young adults, M = 13.63, SD = 9.90, t (1744.7) = 7.88, p < .001, r = .19, but middle-aged adults were not slower than young adults, t (1778) = 0.07, p = .95, r = .002. A second one-way ANOVA showed that the length of time (s) for subjects to provide a confidence rating was also dependent on age group, F (2, 2667) = 45.40, p < .001. Older adults, M = 9.73, SD = 6.17, were slower than both middle-aged, M = 7.88, SD = 6.35, t (1776.5) = 6.23, p < .001, r = .15, and young adults, M = 7.13, SD = 5.16, t (1723.7) = 9.63, p < .001, r = .23. Middle-aged adults were also slower than young adults, t (1705.7) = 2.72, p = .007, r = .07. Together, these analyses suggest that our older adults showed the speed deficits in performance that accompany normal ageing, despite being sampled from a pool of educationally active older adults.

Fair lineups. Before collapsing across the three fair lineup techniques (replication, pixelation and block) in our dataset, we checked that, within each age group, subjects performed similarly on the three fair lineup types. The identification responses made by the young, middle-aged and older adults in the replication, pixelation, block and do-nothing lineups, are presented in Figure C.1 in Appendix C. Three 3 (lineup type: replication, pixelation, block) × 3 (identification response: guilty suspect, foil, rejection) two-way chi-square analyses indicated that performance was the same on the three fair lineups in the young, χ^2 (4, N = 688) = 2.25, p = .69, middle-aged, χ^2 (4, N = 688) = 1.90, p = .75, and older, χ^2 (4, N = 688) = 7.37, p = .12, adults. ROC analyses and fitting a signal detection process model also corroborated that, within each age group, performance on the fair lineups was

similar (see Appendix C). Therefore, for ease of interpretation, we collapsed the data over the replication, pixelation and block lineups within each age group.

Identification responses

We first analysed our data in a way that is consistent with much of the existing eyewitness identification and ageing literature. We calculated the proportion of suspect identifications, foil identifications and lineup rejections (i.e., "Not Present" responses) in the fair and unfair lineups. Figure 3.1 shows the identification responses made by the young, middle-aged and older adults in (a) target-present and (b) target-absent lineups, as a function of lineup type. For target-absent lineups, we calculated the number of innocent suspect and foil identifications in the same way as in Chapter 2.

Target-present lineups. Figure 3.1a shows that there was a decline in the number of accurate responses with age. We conducted a 3 (age: young, middle-aged, older) × 2 (lineup type: fair, unfair) × 3 (identification response: guilty suspect, foil, incorrect rejection) hierarchical loglinear analysis. There was a significant two-way interaction, indicating that age influenced identification responses, χ^2 (4, N = 1,359) = 30.82, $p < .001$ (likelihood ratio: χ^2 (8) = 37.10, $p < .001$). Although all three age groups made a similar number of lineup rejections, the number of guilty suspect IDs decreased and the number of foil IDs increased with age. Older adults made fewer guilty suspect IDs ($z = –2.98$, $p < .01$) but more foil IDs ($z = 2.59$, $p < .01$) than expected, and young adults made more guilty suspect IDs ($z = 2.29$, $p < .05$) and fewer foil IDs ($z = –1.95$, $p > .05$) than expected. Three 2 (age) × 2 (identification response: guilty suspect, foil) two-way chi-square analyses indicated that when subjects made a selection from the lineup, older adults were 1.71 times more likely to identify a foil than middle-aged adults, χ^2 (1, N = 668) = 11.76, $p < .001$, OR = 1.71, 95% CI [1.24, 2.37], and 2.15 times more likely to identify a foil than young adults, χ^2 (1, N = 676) = 23.42, $p < .001$, OR = 2.15, 95% CI [1.56, 2.99]. But middle-aged adults were not significantly more likely to identify a foil than young adults, χ^2 (1, N = 692) = 2.03, $p = .15$, OR = 1.26, 95% CI [0.91, 1.75]. In short, older subjects made more incorrect IDs and fewer correct IDs in target-present lineups than their middle-aged and young counterparts.

The loglinear analysis also revealed a significant two-way interaction indicating that lineup technique influenced identification responses, χ^2 (4, N = 1,356)

= 112.05, $p < .001$ (likelihood ratio: χ^2 (6) = 118.33, $p < .001$). Fair lineups led to fewer guilty suspect IDs ($z = -3.76$, $p < .001$), but more foil IDs ($z = 3.04$, $p < .01$) and more rejections ($z = 1.82$, $p > .05$) than expected. Conversely, unfair lineups led to more guilty suspect IDs ($z = 6.52$, $p < .001$), but fewer foil IDs ($z = -5.27$, $p < .001$) and fewer rejections ($z = -3.15$, $p < .01$) than expected. Specifically, subjects were 3.80 times more likely to make a correct identification in the unfair lineups compared to the fair lineups, χ^2 (1, N = 1,356) = 104.72, $p < .001$, OR = 3.80, 95% CI [2.90, 5.00]. This suggests that when the guilty suspect stood out in target-present lineups, there was an increase in guilty suspect IDs along with a reduction in both foil IDs and incorrect rejections in all age groups.

Target-absent lineups. Figure 3.1b shows that there was a decline in the number of accurate (reject) responses with age. We conducted a 3 (age: young, middle-aged, older) × 2 (lineup type: fair, unfair) × 3 (identification response: innocent suspect, foil, correct rejection) hierarchical loglinear analysis. The two-way interaction between age and identification response did not reach statistical significance, χ^2 (4, N = 1,311) = 7.36, $p = .11$ (likelihood ratio: χ^2 (8) = 14.31, $p = .07$), but the numerical trends indicated that the number of lineup rejections decreased and the number of foil IDs increased with age. Three 2 (age) × 2 (identification response: incorrect, correct rejection) two-way chi-square analyses indicated that older adults were 1.38 times more likely to make an incorrect identification than young adults, χ^2 (1, N = 874) = 5.32, $p = .02$, OR = 1.38, 95% CI [1.04, 1.84], but not significantly more likely to make an incorrect identification than middle-aged adults, χ^2 (1, N = 874) = 2.40, $p = .12$, OR = 1.24, 95% CI [0.93, 1.66]. Middle-aged adults were not significantly more likely to make an incorrect identification than young adults, χ^2 (1, N = 874) = 0.58, $p = .45$, OR = 1.11, 95% CI [0.84, 1.47]. In short, older subjects made more incorrect IDs in target-absent lineups than their young counterparts.

The loglinear analysis also revealed a significant two-way interaction indicating that lineup type influenced identification responses, χ^2 (4, N = 1,311) = 155.01, $p < .001$ (likelihood ratio: χ^2 (6) = 162.37, $p < .001$). Although fair and unfair lineups led to a similar number of lineup rejections (fair: $z = 0.71$, $p > .05$; unfair: $z = -1.40$, $p > .05$), fair lineups led to fewer innocent suspect IDs ($z = -5.38$, $p < .001$), but more foil IDs ($z = 2.65$, $p < .01$) than expected. Conversely, unfair

lineups led to more innocent suspect IDs ($z = 10.63$, $p < .001$), but fewer foil IDs ($z = –5.25$, $p < .001$) than expected. Specifically, when subjects made an identification, they were 13.96 times more likely to identify the innocent suspect in the unfair lineups compared to the fair lineups, χ^2 (1, $N = 875$) = 259.45, $p < .001$, OR = 13.96, 95% CI [9.69, 20.34]. This suggests that when the innocent suspect stood out in target-absent lineups, subjects in all age groups shifted their identifications from the other lineup members onto the innocent suspect.

In sum, these results are concordant with the existing literature and indicate that the number of erroneous identifications increased with age. Unfair lineups also led to more correct identifications in target-present lineups but more incorrect identifications of innocent suspects in target-absent lineups, in all age groups.

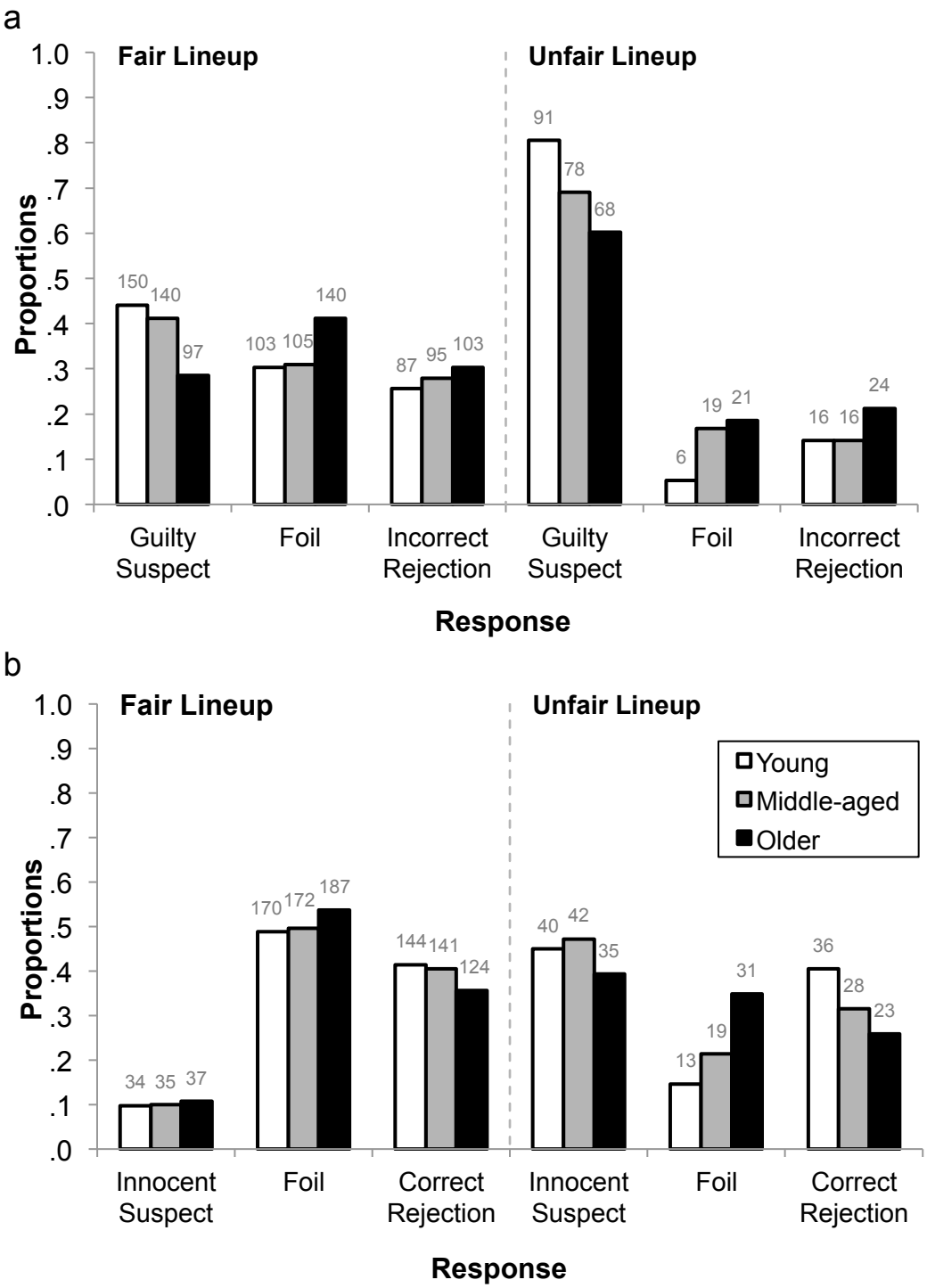

Figure 3.1. Identification responses made by the young, middle-aged, and older adults in fair and unfair (a) target-present and (b) target-absent lineups. Data labels are absolute frequencies.

ROC analysis

Next, we conducted ROC analysis to investigate whether the patterns in our identification responses were due to changes in subjects' ability to discriminate between guilty and innocent suspects, or subjects' willingness to identify the suspect. We constructed our ROC curves and calculated our *p*AUC statistics in the same way as in Chapter 2. We set the specificity (1 – FAR) using the FAR range covered by the least extensive curve to .902. Figure 3.2 shows the ROC curves for the fair and unfair lineups in the young, middle-aged and older subjects.

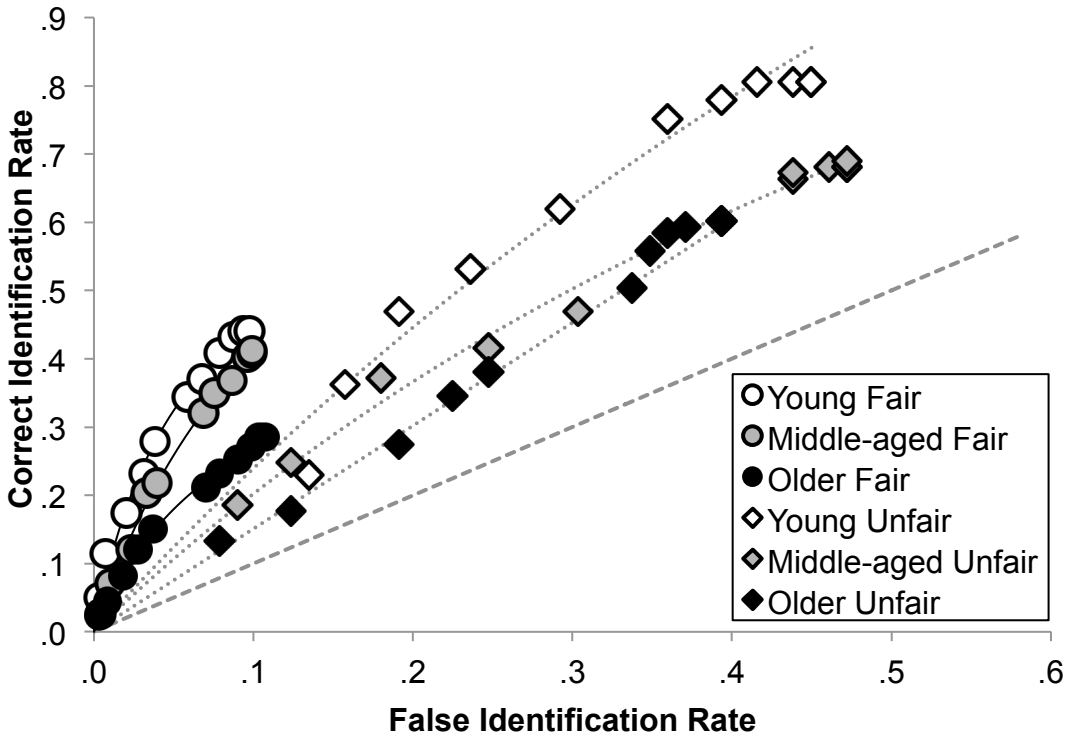

Figure 3.2. Receiver operating characteristic (ROC) curves for the fair and unfair lineups for the young, middle-aged, and older adults. The dashed line represents chance-level performance.

Fair lineups. Considering the fair lineups in Figure 3.2, it is evident that ability to discriminate between innocent and guilty suspects declined with age. The *p*AUC for the older adults (*p*AUC = 0.016, 95% CI: 0.011, 0.021) was, descriptively speaking, smaller than the *p*AUC for the middle-aged adults (*p*AUC = 0.024, 95% CI: 0.017, 0.031, $D = 1.65$, $p = .10$), and significantly smaller than the *p*AUC for the young adults (*p*AUC = 0.028, 95% CI: 0.022, 0.036, $D = 2.68$, $p = .007$). The *p*AUC

for the middle-aged adults was also smaller than the *p*AUC for the young adults, but not significantly so ($D = 0.92$ $p = .36$).

***Unfair lineups*.** Considering the unfair lineups in Figure 3.2, however, the ROC curves for each age group are largely overlapping and close to the dashed chance line. This indicates that the ability to discriminate between innocent and guilty suspects in unfair lineups was similar, and poor, in all age groups. The *p*AUC for the older adults (*p*AUC = 0.008, 95% CI: 0.003, 0.016) was similar to the *p*AUC for both the middle-aged (*p*AUC = 0.010, 95% CI: 0.005, 0.021, $D = 0.37$, $p = .71$) and the young (*p*AUC = 0.008, 95% CI: 0.005, 0.018, $D = 0.05$, $p = .96$) adults. The *p*AUC for the middle-aged adults was also similar to the *p*AUC for the young adults ($D = 0.37$, $p = .71$). All three age groups were less able to distinguish between innocent and guilty suspects in the unfair lineups than in the fair lineups. Indeed, the *p*AUC for the unfair lineups was significantly smaller than the *p*AUC for the fair lineups in the young ($D = 3.94$, $p < .001$), middle-aged ($D = 2.53$, $p = .01$), and older ($D = 1.96$, $p = .05$) adults. Finally, Figure 3.2 shows that the ROC curves for the unfair lineups are shifted to the right of the ROC curves for the fair lineups, reflecting an increase in both correct and false identifications. In line with the identification response data, the ROC results indicate that subjects of all ages were more willing to identify the suspect when the suspect was the only person in the lineup with the distinctive feature.

In sum, the ROC results indicate that the ability to discriminate between guilty and innocent suspects substantially declined with age in fair lineups, but all age groups were poor at sorting guilty and innocent suspects into their appropriate categories in unfair lineups. All subjects were more willing to identify the suspect in the unfair lineups compared to the fair lineups.

Modelling

To further test these conclusions, we fit a signal detection model to our data (Wixted & Mickes, 2014). The pattern of results found in our model fitting aligned with the results of our ROC analyses (see Appendix D), which indicates that the findings of our atheoretical *p*AUC analysis map onto measures of underlying memory discriminability (cf. Lampinen, 2016). Here we limit our discussion to our findings when we fit the model to the fair lineups because this also furthers our theoretical understanding of how identification behaviour changes with age. The

model accounts for all identification decisions (suspect identifications, foil identifications and lineup rejections in both target-present and target-absent lineups). Therefore, the model fitting is theoretically valuable because it helps us to understand the decision-making processes of witnesses and illustrates how willingness to make identifications (i.e., placement of the decision criterion) changes with differences in discriminability (see Palmer & Brewer, 2012 for a discussion).

The model assumes that guilty suspects, innocent suspects and foils each have memory strength values with Gaussian distributions and means of μ_{guilty}, $\mu_{innocent}$, and μ_{foil}, respectively. In a fair lineup, $\mu_{innocent} = \mu_{foil}$, therefore the model consists of two distributions: one for guilty suspects (μ_{guilty}), and one for innocent suspects and foils ($\mu_{innocent}$). The distance between the μ_{guilty} and $\mu_{innocent}$ distributions (d') measures subjects' underlying ability to discriminate between who is guilty and who is innocent. Smaller values of d' reflect poorer discriminability (see Chapter 1 for a description of the model).

The model also assumes that there is a set of response criteria that reflect different levels of confidence. To limit the number of parameters, we collapsed our data from the 11-point confidence scale used in the ROC analysis (0, 10, 20, etc.), down to a 5-point confidence scale: 0–20 (c_1), 30–40 (c_2), 50–60 (c_3), 70–80 (c_4), and 90–100 (c_5). We used these confidence intervals for two main reasons: (1) they ensured a relatively similar number of identification decisions at each confidence level in each age group and lineup type, and (2) they ensured consistency throughout our analyses because we used these intervals when constructing our confidence-accuracy plots to provide more stable estimates in each confidence category (following, for instance, Brewer & Wells, 2006; Sauer et al., 2010). The model assumes that the lineup is rejected if no face is familiar enough to exceed the lowest decision criterion (c_1). Conversely, an identification is made when the familiarity of one or more faces exceeds c_1, and the face which is identified is simply the face with the highest familiarity value. The confidence in the identification is determined by the highest criterion that is exceeded.

If the increase in erroneous identifications made by the middle-aged and older adults is due to impairment in underlying theoretical discriminability, then there should be a greater overlap of the guilty and innocent distributions (i.e., d' should decline) with age. However, if the increase in erroneous identifications made by the

middle-aged and older adults is due to more liberal responding, then there should be a marked leftward shift of the decision criteria (i.e., c_1 through c_5 should all decline) with age. The data contained 15 degrees of freedom, corresponding to the 5 levels of confidence for guilty suspect identifications and foil identifications in target-present lineups, and the 5 levels of confidence for foil identifications in target-absent lineups. Once these response frequencies were known, the number of rejections made in target-present and target-absent lineups was fixed. The model had 6 free parameters (μ_{guilty}, c_1, c_2, c_3, c_4, c_5) because we fixed $\mu_{innocent}$ to 0 and set the standard deviations for each distribution to 1, for simplicity. Thus, the fit had 15 – 6 = 9 degrees of freedom.

We fit the model to our young, middle-aged and older adults' data by minimising the chi-square goodness-of-fit statistic. Table 3.2 shows our observed data and the values predicted by the best-fitting model for each age group, while Table 3.3 shows the best-fitting parameters and the chi-square goodness-of-fit statistics. It is clear from Table 3.2 that the model proficiently captured the trends in our data, and this is reflected in the (non-significant) chi-square goodness-of-fit statistics in the left-hand column (full model) of Table 3.3. Non-significant chi-square goodness-of-fit statistics (i.e., $p > .05$) indicate that the data do not significantly deviate from the model-predicted values, that is, they indicate that the model fits the data well. Notably, Figure 3.3 displays the parameters estimated by the best-fitting model for the three age groups. Evidently, the overlap in the guilty and innocent distributions increases (i.e., d' declines) with age. Interestingly, the decision criteria also spread out on the decision axis from young to older subjects. Perhaps this trend is more easily observed by looking at the confidence parameter estimates for the young and older adults displayed numerically in the left-hand column (full model) of Table 3.3. A larger confidence parameter estimate corresponds to a more conservative confidence criterion setting, whereas a smaller confidence parameter estimate corresponds to a more liberal confidence criterion setting. Compared to young adults, older adults set their high-confidence criteria (i.e., c_4 and c_5) in a more conservative position, but place their remaining criteria (i.e., c_1, c_2, c_3) in a more liberal position. A similar pattern has been observed when memory strength is manipulated in studies of younger subjects and is a natural consequence of a decline in d' (Stretch & Wixted, 1998).

Table 3.2

Observed and Predicted Identification Responses in Each Confidence Bin in the Fair Lineups for the Young, Middle-aged, and Older Adults

	Target present			Target absent	
Confidence	Guilty suspect	Foil	Incorrect rejection	Foil	Correct rejection
			Young		
0–20					
Observed	11.00	19.00	-	40.00	-
Predicted	13.26	17.88	-	38.66	-
30–40					
Observed	22.00	28.00	-	42.00	-
Predicted	20.58	24.22	-	46.93	-
50–60					
Observed	38.00	25.00	-	57.00	-
Predicted	33.45	31.72	-	53.97	-
70–80					
Observed	40.00	21.00	-	50.00	-
Predicted	42.57	27.94	-	41.29	-
90–100					
Observed	39.00	10.00	-	15.00	-
Predicted	36.72	11.61	-	15.23	-
Total					
Observed	-	-	87.00	-	144.00
Predicted	-	-	80.04	-	151.92
			Middle-aged		
0–20					
Observed	15.00	21.00	-	26.00	-
Predicted	11.38	17.40	-	34.28	-
30–40					
Observed	16.00	13.00	-	38.00	-
Predicted	13.56	18.86	-	34.45	-
50–60					
Observed	40.00	35.00	-	74.00	-
Predicted	36.40	41.86	-	68.42	-
70–80					
Observed	45.00	21.00	-	48.00	-
Predicted	37.96	30.75	-	44.01	-
90–100					
Observed	24.00	15.00	-	21.00	-
Predicted	30.94	13.14	-	17.01	-
Total					
Observed	-	-	95.00	-	141.00
Predicted	-	-	87.73	-	149.83
			Older		
0–20					
Observed	11.00	23.00	-	35.00	-
Predicted	10.22	22.20	-	36.12	-
30–40					
Observed	14.00	25.00	-	42.00	-
Predicted	13.11	26.42	-	40.80	-
50–60					
Observed	31.00	61.00	-	89.00	-
Predicted	35.04	59.86	-	85.31	-
70–80					
Observed	26.00	25.00	-	40.00	-
Predicted	22.93	29.49	-	38.70	-
90–100					
Observed	15.00	6.00	-	18.00	-
Predicted	13.99	11.62	-	14.56	-
Total					
Observed	-	-	103.00	-	124.00
Predicted	-	-	95.12	-	132.49

Note. The total row displays all reject identification decisions because the model does not account for the confidence level with which lineup rejections are made.

Table 3.3
Full and Constrained (d') Model Fits for the Young vs. Middle-aged, Young vs. Older, and Middle-aged vs. Older Fair Lineup Comparisons

	Full model		Constrained model	
Estimate	Young	Middle-aged	Young	Middle-aged
μ_{guilty} (d')	1.21	1.07	1.14	1.14
c_1	1.13	1.12	1.12	1.13
c_2	1.31	1.28	1.30	1.29
c_3	1.54	1.44	1.53	1.46
c_4	1.89	1.86	1.88	1.87
c_5	2.44	2.39	2.42	2.41
Overall χ^2	26.12		27.99	
Overall df	18		19	
Overall p	.10		.08	
	Young	Older	Young	Older
μ_{guilty} (d')	1.21	0.72	0.99	0.99
c_1	1.13	1.04	1.09	1.08
c_2	1.31	1.21	1.27	1.25
c_3	1.54	1.40	1.49	1.44
c_4	1.89	1.92	1.84	1.97
c_5	2.44	2.45	2.36	2.50
Overall χ^2	15.90		38.33	
Overall df	18		19	
Overall p	.60		.005	
	Middle-aged	Older	Middle-aged	Older
μ_{guilty} (d')	1.07	0.72	0.91	0.91
c_1	1.12	1.04	1.07	1.09
c_2	1.28	1.21	1.23	1.25
c_3	1.44	1.40	1.42	1.41
c_4	1.86	1.92	1.96	1.83
c_5	2.39	2.45	2.49	2.37
Overall χ^2	23.98		35.29	
Overall df	18		19	
Overall p	.16		.01	

Note. The full model allows d' to differ between the two age groups being compared. The constrained model holds d' constant across the two age groups being compared. Overall χ^2, df and p rows represent goodness-of-fit statistics when the model was fit to the two age groups together.

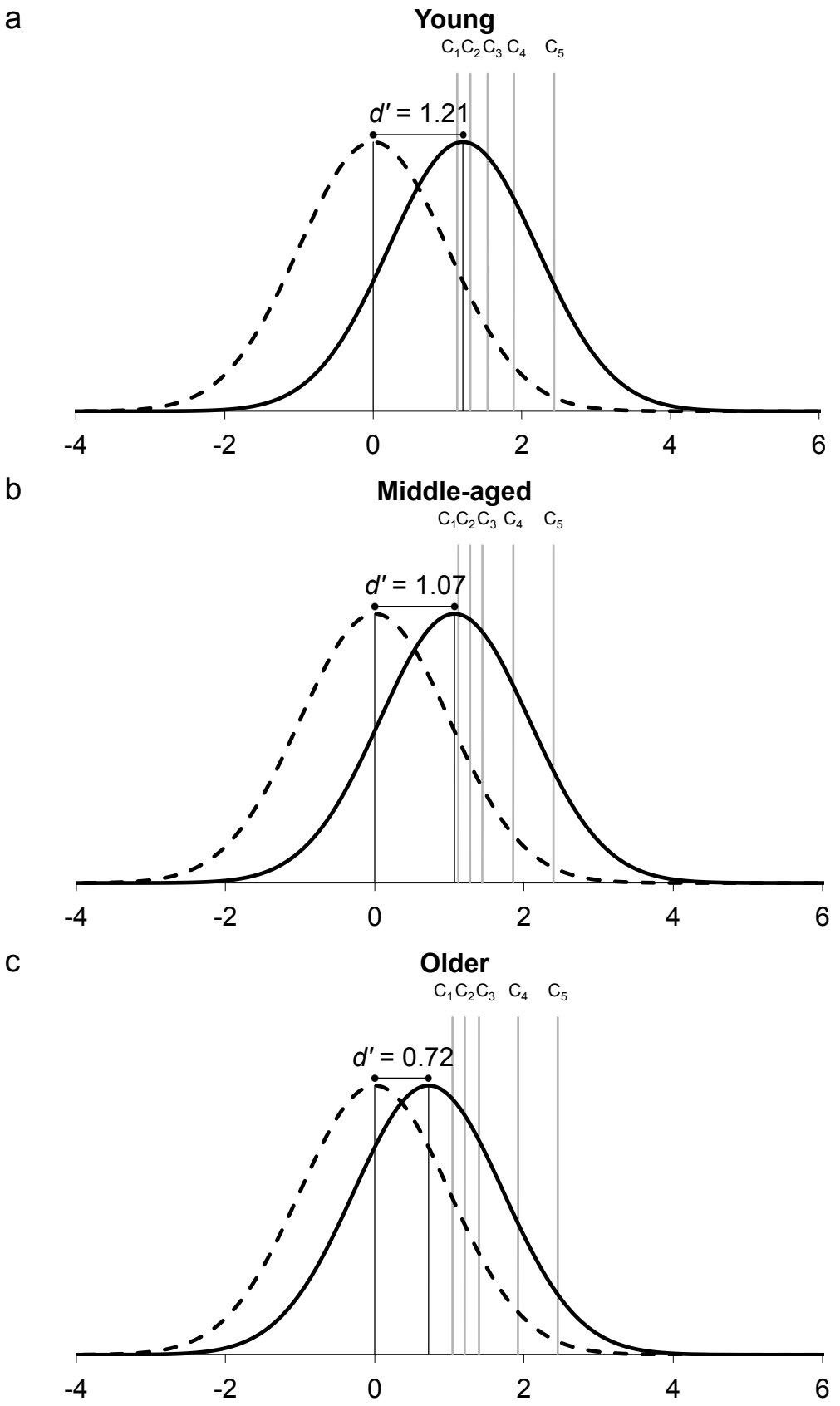

Figure 3.3. Innocent and guilty distributions for (a) young, (b) middle-aged, and (c) older adults using the best-fitting signal detection model parameters. *d′* measures subjects' ability to discriminate between innocent and guilty faces. c_1, c_2, c_3, c_4 and c_5 are a set of response criteria that reflect different levels of confidence.

Discriminability. To test whether the observed decline in d' with age was statistically significant, we performed three pairwise comparisons: young versus middle-aged, young versus older, and middle-aged versus older. We fit the same model, allowing the confidence criteria to differ, but we constrained d' to be equal in the two age groups being compared. The overall χ^2, df and p rows in Table 3.3 show the full (unconstrained) and constrained model fit statistics. In comparison to the full model, the constrained model did not provide a significantly worse fit of the data for the young and middle-aged comparison, χ^2 (1) = 1.87, p = .17, but it did provide a significantly worse fit of the data for the young and older, χ^2 (1) = 22.43, $p < .001$, and middle-aged and older, χ^2 (1) = 11.31, $p < .001$, comparisons. These results indicate that ageing is accompanied by a decline in theoretical discriminability, but the decline from young adults to middle-aged adults was not statistically significant.

Decision criteria. To examine the manner in which the decision criteria changed with age, we tested the difference in criteria settings in the young versus older adults. Figure 3.4 shows the best-fitting model confidence criteria parameters for the young versus older adults. The confidence criteria for the young and older adults are linearly related; therefore we fit the same model, but we replaced the 5 confidence parameters (c_1, c_2, c_3, c_4, c_5) for the older adults with a linear transformation of the 5 confidence parameters for the young adults. For instance, $c_{1old} = a * c_{1young} + b$, where a and b are free parameters. We allowed d' to differ across the young and older groups. The overall χ^2, df and p rows in Table 3.4 show the full (unconstrained confidence parameters) and reduced (linear transformation of c_1 - c_5) model fit statistics. The model fit statistic in Table 3.4 indicates that the reduced (linear transformation of c_1 - c_5) model fit the data well, but, surprisingly, it provided a significantly worse fit of the data than the full model, χ^2 (3) = 12.70, p = .01. Looking back at Figure 3.4, it is clear that c_3 falls slightly away from the line of best fit. Therefore, it is possible that this one criterion could explain why the fit of the reduced (linear transformation of c_1 - c_5) model was significantly worse than the fit for the full model. To address this, we fit the same linear transformation model, but this time we allowed c_3 to vary across the young and old groups. The overall χ^2, df and p rows in Table 3.4 show the reduced (linear transformation of c_1, c_2, c_4, c_5) model fit statistic and indicate that the model fit the data well. This time, the new reduced (linear transformation of c_1, c_2, c_4, c_5) model did not provide a significantly

worse fit of the data than the full model, χ^2 (2) = 3.02, p = .22. This suggests that a linear transformation, while allowing c_3 to vary, adequately characterises the confidence criteria in the young versus older groups.

Next, we fit the same model, but this time we equated the confidence parameters in the young and older groups, setting a = 1 and b = 0. Again, we allowed d' to differ across the young and older groups. The overall χ^2, df and p rows in Table 3.4 show the reduced (linear transformation of c_1, c_2, c_4, c_5) and constrained (equated confidence parameters) model fit statistics. In comparison to the reduced model, the constrained model provided a significantly worse fit of the data, χ^2 (3) = 15.52, p = .001. These results indicate that ageing is accompanied by a statistically significant change in criteria settings. This change is, generally speaking, linear, suggesting that the older adults tend to spread out their decision criteria more than the young adults. Setting the high-confidence criterion to a more conservative position, while spreading the remaining decision criteria to more liberal positions in this way at least approximates an optimal strategy because it means that identifications made with high confidence are likely to remain highly accurate, even though there is a general decline in d' (Stretch & Wixted, 1998). Thus, this provides preliminary evidence that older adults adjust their criteria in a way that maintains a good confidence-accuracy relationship.

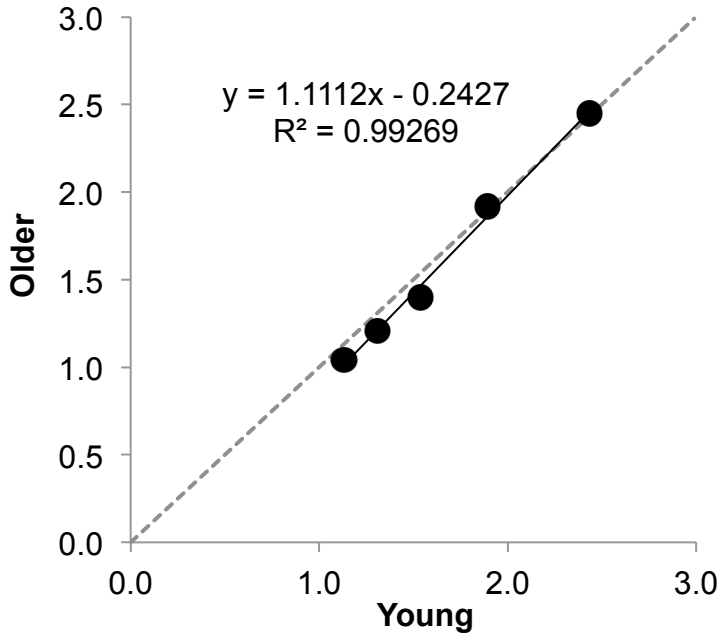

Figure 3.4. The best-fitting signal detection model confidence criteria parameters (c_1, c_2, c_3, c_4, c_5) for the young vs. older adults. The dashed line is y = x.

Table 3.4
Full, Reduced, and Constrained (Confidence Criteria) Model Fits for the Young vs. Older Fair Lineup Comparisons

	Full model		Reduced (linear, c_1 - c_5) model		Reduced (linear, c_1, c_2, c_4, c_5) model		Constrained model	
Estimate	Young	Older	Young	Older	Young	Older	Young	Older
μ_{guilty} (d')	1.21	0.72	1.20	0.71	1.20	0.72	1.17	0.76
c_1	1.13	1.04	1.13	1.03	1.13	1.04	1.08	1.08
c_2	1.31	1.21	1.30	1.21	1.29	1.22	1.25	1.25
c_3	1.54	1.40	1.50	1.43	1.54	1.40	1.46	1.46
c_4	1.89	1.92	1.92	1.89	1.91	1.90	1.91	1.91
c_5	2.44	2.45	2.44	2.45	2.42	2.47	2.44	2.44
a	-		1.09		1.11		1.00	
b	-		–0.21		–0.21		0.00	
Overall χ^2	15.90		28.60		18.92		34.44	
Overall df	18		21		20		23	
Overall p	.60		.12		.53		.06	

Note. The full model allows the confidence criteria (c_1 - c_5) to differ between the young and older groups. The reduced (linear, c_1 - c_5) model allows the confidence criteria to differ between the young and older groups by a linear transformation. The reduced (linear, c_1, c_2, c_4, c_5) model allows the confidence criteria c_1, c_2, c_4, and c_5 to differ between the young and older groups by a linear transformation, and leaves c_3 free to vary. The constrained model holds the confidence criteria constant across the young and older groups. Overall χ^2, df and p rows represent goodness-of-fit statistics when the model was fit to the two age groups together.

Confidence and accuracy

So far, our analyses have illustrated that when subjects are presented with a fair lineup, discriminability declines with age, but older adults spread out their decision criteria in a more-or-less optimal manner. This result seems to indicate that middle-aged and older adults are aware that their memory accuracy is poor and that they make adjustments accordingly. Here, we tested this idea more concretely. If middle-aged and older subjects realise that their memory is error-prone, they should lower their confidence judgements to reflect their poorer performance and the proportion of correct identifications should be similar in all three age groups at each level of confidence.

To test this, we constructed confidence-accuracy curves in the same way as in Chapter 2. The frequencies of identification responses in each confidence bin are

presented in Appendix E. Figure 3.5 shows the confidence-accuracy curves for fair and unfair lineups in the young, middle-aged and older subjects. For fair lineups, the error bars for each age group largely overlap. This indicates that the differences in suspect identification accuracy between the three age groups at each level of confidence are not, on the whole, statistically reliable (e.g., Sauer et al., 2010). Despite being significantly poorer at distinguishing between who is guilty and who is innocent than the young and middle-aged adults, older adults seem to be reasonably effective at regulating their confidence judgements to reflect the likely accuracy of their suspect identification decisions. This means that they are aware that their memory is poorer and adjust their confidence criteria accordingly. Nevertheless, descriptively speaking, Figure 3.5 shows that the older adults are slightly less accurate at every level of confidence than the young and middle-aged adults. This suggests that older adults, while they do adjust their confidence criteria in the appropriate direction, do not quite adjust their confidence criteria enough, given their decline in memory ability. For example, if we look back at the signal detection model parameters illustrated in Figure 3.3, older adults would need to set c_5 to a more conservative position if they were to be as accurate as the young and middle-aged adults at the highest level of confidence (i.e., 90–100 certain).

Finally, comparing the fair and unfair lineups in Figure 3.5, suspect identification accuracy is reduced in the unfair lineups in all age groups. Specifically, within each age group, high-confidence suspect identifications made in unfair lineups were substantially less trustworthy than high-confidence suspect identifications made in fair lineups. This suggests that subjects were not aware that their accuracy was poor in the unfair lineups and did not adjust their confidence accordingly.

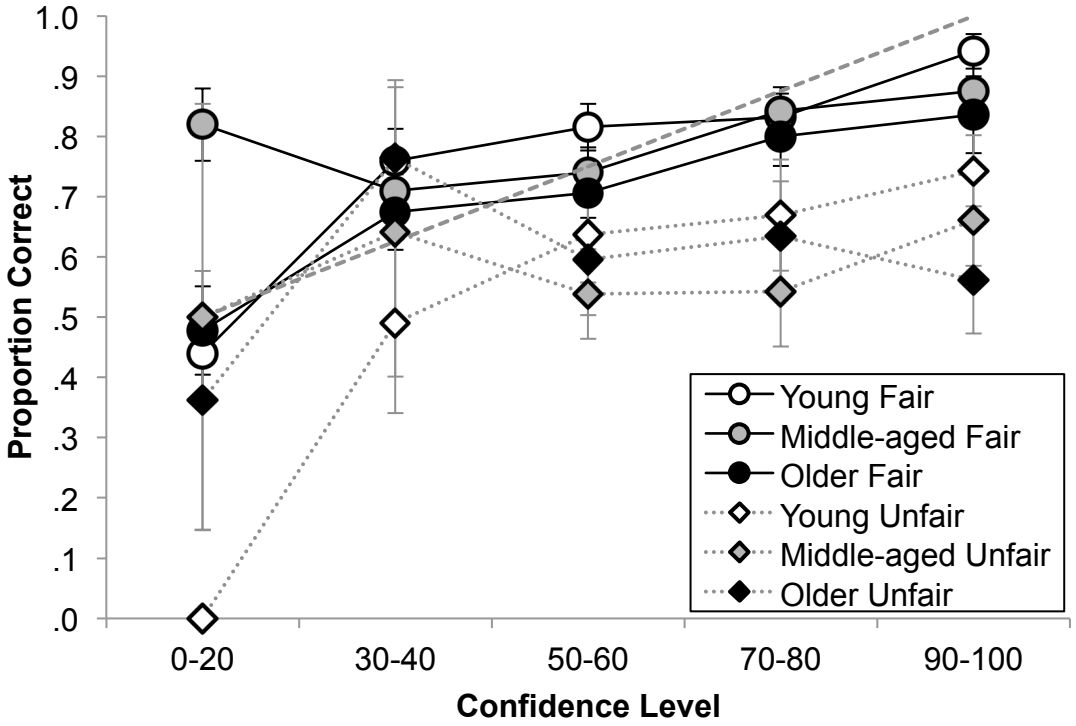

Figure 3.5. Confidence-accuracy curves for suspect identifications in the fair and unfair lineups. Error bars indicate ±1 SE. The dashed line represents chance accuracy at the lowest confidence bin (i.e., 0–20) and perfect accuracy at the highest confidence bin (i.e., 90–100).

Discussion

We asked [1] whether the age-related decline in accurate identification decisions is due to an increased willingness to make an identification, a decline in discriminability, or both, and [2] whether older and middle-aged adults are able to gauge the likely accuracy of their suspect identification decisions and assign appropriate confidence judgements to the same extent as young adults. Our findings suggest that ageing is associated with a genuine decline in ability to discriminate between who is innocent and who is guilty. Remarkably, despite a substantial decline in memory ability, older adults were able to gauge the accuracy of their suspect identifications, and were, generally speaking, as accurate as the young and middle-aged adults at each level of confidence.

At first glance, our results are perhaps unsurprising. Many previous studies have shown that older adults make more mistakes on lineup tasks than younger adults (see Bartlett & Memon, 2007 and Sporer & Martschuk, 2014 for reviews). Indeed, the distribution of identification responses indicated that the number of

erroneous identifications increased with age. But our analyses show that this pattern of results is not simply due to older adults being more willing to make an identification from a lineup. Instead, our data suggest that the errors are due to a genuine decline in ability to discriminate between those who are innocent and guilty.

Why might ageing be associated with a genuine decline in recognition performance? One explanation is that our ability to recollect source-specific information declines over the lifespan, which results in a greater reliance on familiarity processes with age (Healy et al., 2005; Searcy et al., 1999). Older adults were more likely to make erroneous identifications in both target-present and target-absent lineups than young adults. Presumably this is because the faces in the lineups were very similar and so even the new faces evoked signals of perceived familiarity (Bartlett et al., 1984; Edmonds et al., 2012; Young et al., 1985). Further support for this theoretical account comes from our model fitting. If older adults are more reliant on a general feeling of familiarity, then the strength of the memory signal from new faces in the lineup (i.e., those in the innocent distribution) should be closer to the strength of the memory signal from the real culprit (i.e., those in the guilty distribution). Indeed, we found a statistically significant increase in the overlap of the innocent and guilty distributions with age.

Our finding that ability to discriminate between innocent and guilty suspects declines with age is concordant with face recognition studies in the broader literature (e.g., Lamont et al., 2005) and the proposition that ageing might be associated with a decline in configural or holistic processing (see Boutet, Taler, & Collin, 2015 for a review), because we know that face recognition is dependent on processing the spatial distances between facial features (configural processing) and processing the face as a whole (holistic processing; see Tanaka & Gordon, 2011 for a review). More specifically, our finding is consistent with a recent meta-analysis in the eyewitness identification domain (Fitzgerald & Price, 2015). Key et al. (2015), by contrast, found equivalent performance in their young and older subjects using ROC analysis. One possible reason for these contradictory findings is that Key et al.'s young and older groups consisted of subjects aged 18–59 and over 60, respectively. Our results suggest that discriminability begins to decline from early adulthood (aged 18–30) to middle age (aged 31–59). Performance in Key et al.'s young group may have been artificially low because of its wide age range. Therefore, it is possible that the non-

significant difference in discriminability between the young and older adults reflected how their young and older age groups were defined.

So, why is all this important? Greater theoretical understanding of how memory changes with healthy ageing can be used to advance appropriate procedures to help aid identification accuracy. Many studies have attempted to reduce older adults' false identification rate by reducing their proclivity to choose (e.g., Memon & Gabbert, 2003; Rose et al., 2005; Wilcock et al., 2005). But our data suggest that encouraging older adults to be more conservative when they make a decision will not reduce the age-related deficit in performance. Instead, our results indicate that procedures need to target middle-aged and older adults' ability to discriminate between who is innocent and who is guilty if the identification errors made by these older age groups are to be reduced.

One might argue that older adults made more identification errors simply because their eyesight was poorer than the young and middle-aged subjects. However, there are at least three reasons why poorer vision in older adults is unlikely to explain our results. First, the older adults, like the young and middle-aged adults, were more willing to identify the suspect in the unfair lineups than in the fair lineups. This suggests that older subjects saw the distinctive feature in the video because they subsequently picked the only person with a distinctive feature during the lineup task. Second, we asked a separate group of young (n = 20, aged 18–30) and older (n = 29, aged 60–85) adults to watch the mock crime video and then describe the culprit's appearance. The proportion of young and older adults who correctly described the distinctive feature did not differ for either the mugging or the graffiti video (ps > .19). This suggests that the vision of both young and older adults was good enough to see and encode the face of the culprit. Finally, the findings from our identification responses analyses are consistent with many laboratory-based studies that likely had greater control over whether subjects were wearing glasses, if necessary (e.g., Badham et al., 2013). Therefore, it seems that recognition memory ability on lineup tasks declines with age.

Perhaps most strikingly, our study has shown that despite older adults' poorer recognition memory ability and speed deficits, suspect identifications made by older adults can be almost as accurate as those made by young and middle-aged adults, when the confidence judgement expressed immediately after the identification

decision is taken into account. In practice, this finding is important for legal decision makers because it means that an identification made with a particular level of confidence is likely to be similarly accurate regardless of whether it is made by a young, middle-aged or older adult. Recall that in our modelling (which accounted for all identification decisions) we found that the confidence criteria naturally spread out along the decision axis as d' declined with age. The fact that there were no significant differences in suspect identification accuracy between the age groups at each level of confidence indicates that the extent of spreading was generally appropriate to account for the decline in d'. Theoretically, this illustrates that older adults are, on the whole, able to assess the likely accuracy of their memories.

Recall also, however, that there was a trend for the older adults to be slightly (but not significantly) less accurate at every level of confidence than the young and middle-aged adults in our confidence-accuracy plot. To investigate this further, we separated our older adults into young-old (aged 60–70) and old-old (aged 71+) groups, and we saw the same numerical pattern: old-old adults were slightly (but not significantly) less accurate at every level of confidence than the young-old adults (see Appendix F). This trend accords with other research that shows that older adults can have reduced metacognitive monitoring of recently encountered information (e.g., Dodson & Krueger, 2006), can experience high-confidence false memories (e.g., Dodson, Bawa, & Krueger, 2007), and sometimes have a tendency for less flexible criterion placement in difficult memory tasks (e.g., Koutstaal, 2006). Thus, there is some basis for the idea that ageing may be associated with a difference in adjusting criteria to account for poorer memory ability. One theory suggests that people are usually adept at assigning appropriate confidence judgements because they have learned through error feedback training the situations in which their memory is and is not likely to be accurate (Mickes, Hwe, Wais, & Wixted, 2011; Stretch & Wixted, 1998; see also D. S. Lindsay et al., 1998). Therefore, it is possible that, as we age, memory ability declines quicker than we are able to learn about the degree of our memory impairment though error feedback training. This might explain why our older adults failed to adjust their confidence criteria to the extent required for them to be just as accurate as the young and middle-aged adults. Notably, this idea is based on trends, and not statistically significant differences, in our data. Therefore, our main conclusion still stands: suspect identifications made by

older adults are as accurate as those made by young and middle-aged adults when their confidence judgement is taken into account. Nevertheless, examining the role of error feedback training in older adults could be a fruitful avenue for further research.

Finally, our comparison between performance on fair and unfair lineups is also important. We found that subjects of all ages were more willing to identify the suspect, but, critically, were also less able tell the difference between innocent and guilty suspects in unfair lineups compared to fair lineups. Indeed, ability to discriminate between innocent and guilty suspects on the unfair lineups was remarkably poor in all age groups. Suspect identification accuracy was also reduced at almost every level of confidence in the unfair lineups, compared to the fair lineups. This suggests that subjects were not aware that their accuracy was poor in the unfair lineups and did not adjust their confidence judgements accordingly. These results replicate the findings from Chapter 2 and reiterate the need for fair lineups for witnesses of all ages. Interestingly, these results are predicted by the diagnostic-feature-detection model, which suggests that witnesses are less able to distinguish between innocent and guilty suspects when they rely on features that both innocent and guilty suspects share (Wixted & Mickes, 2014). Our fair lineups prevented subjects from relying on the distinctive feature to make their identification decision because the feature was either concealed (pixelation and block) or appeared on every lineup member (replication). Our unfair lineups, however, did not provide this protection because only one lineup member—the suspect—had the distinctive feature. According to the diagnostic-feature-detection account, subjects who viewed our unfair lineups relied on the distinctive feature to make their identification decision and this impaired their ability to tell the difference between innocent and guilty suspects because the feature was something that both the innocent and guilty suspect shared. Theoretically, then, our research lends support for the idea that fair lineups enhance people's ability to discriminate between innocent and guilty suspects because fair lineups promote reliance on facial features that are diagnostic of guilt, whereas unfair lineups do not.

To conclude, we have shown that errors made by older individuals on lineup tasks are likely attributable to a genuine decline in ability to tell the difference between who is innocent and who is guilty, rather than an increased willingness to

make an identification. Although further research is required before practical recommendations are made to the Criminal Justice System, our results add to the growing literature that suggests that if you were a police officer you should always use fair lineups to enhance your witness's accuracy. But, crucially, our results provide new, preliminary evidence that if you were a judge considering an identification made at a particular level of confidence, you should impart the same amount of trust in the identification regardless of whether it was made by a young, middle-aged or older eyewitness.

Chapter 4 :
Identification Performance and Appearance Change

"Defendant's tattoo did not make the live lineup impermissibly suggestive. None of the witnesses observed a tattoo on the gunman's head."

People v. Gonzalez (2006)

Overview

When constructing lineups for suspects with distinctive features (tattoos, piercings, scars, etc.), police officers must ensure that the suspect does not stand out. So far, our research has shown that, when the culprit has a distinctive feature during the crime, three fair lineup techniques for accommodating distinctive suspects—replication, pixelation and block—are equally effective and all enhance people's ability to discriminate between innocent and guilty suspects compared to unfair (do-nothing) lineups in which the suspect is the only person with the distinctive feature. All three fair lineups are equally effective within different age groups (i.e., young, middle-aged and older adults) and fair lineups enable adults of all ages to assess the likely accuracy of their memories and assign appropriate confidence judgements. Together, these studies provide support for the diagnostic-feature-detection model (Wixted & Mickes, 2014), demonstrate the dangers of unfair lineups and suggest that there are multiple effective routes to create fair lineups for distinctive suspects when the culprit has a distinctive feature at the time of the crime.

But how do the lineup techniques for distinctive suspects influence identification performance when the culprit does not have a distinctive feature at the time of the crime, yet the suspect has a feature at the time of the lineup? In this chapter, we conducted a single experiment (N = 1,463) to compare the fair (replication, pixelation, and block) and unfair (do-nothing) lineup techniques in subjects who had watched a video of a mock crime being committed by a culprit either without a distinctive feature, or by the same culprit with a distinctive feature.

Introduction

When friends get tattoos, we still recognise them with ease, despite changes to their appearance. But what happens when an unfamiliar person, say, a criminal

culprit that we have only seen once, changes their appearance? How well would we recognise that person when they are placed in a police lineup and we are asked to make an identification? Police officers have no control over a culprit's changing appearance, and they cannot possibly know if a culprit's appearance has changed between the crime and the point at which the suspect is put into a lineup. One thing police officers do control, though, is how the lineup is constructed. In this chapter, we examine how different lineup techniques for accommodating suspects with distinctive features affects eyewitness identification behaviour when the culprit does not have the distinctive feature during the crime.

As many as a third of suspects may have a distinctive facial feature, such as tattoos, scars, chipped teeth, or bruising around the eyes (Flowe et al., 2010). Guidelines for constructing lineups for suspects with distinctive features typically suggest that police officers must prevent distinctive suspects from standing out to ensure that every lineup member is a plausible alternative to the suspect (e.g., Police and Criminal Evidence Act 1984, Code D, 2011; Technical Working Group for Eyewitness Evidence, 1999). Preventing distinctive suspects from standing out reduces the chance of witnesses with poor memories simply selecting the suspect because he is obviously the focus of the police investigation or he looks different in some way (Charman, Wells, & Joy, 2011; Wells et al., 1998).

Police officers have several methods for preventing distinctive suspects from standing out. One technique involves digitally replicating the suspect's distinctive feature across the foils (replication). Another technique is to digitally conceal the area of the feature on the suspect's face and to conceal a similar area on the foil faces. Typically, officers will conceal features either by pixelating the same area, or placing a solid black block onto the same area, on each of the lineup faces (pixelation and block, respectively). All three of these techniques create "fair" lineups because they ensure that the foils match the appearance of the suspect.

Recent research has tested the efficacy of the three fair lineup techniques—replication, pixelation and block—by comparing them to unfair lineups in which the suspect is the only person with the distinctive feature (do-nothing lineups). Subjects watched a video of a culprit with a distinctive facial feature committing a mock crime, and then, after a brief delay, attempted to recognise the culprit (i.e., the guilty suspect) from a replication, pixelation, block or do-nothing lineup. The findings

suggest that all three fair lineup techniques are equally effective, and all three fair techniques enhance people's ability to discriminate between innocent and guilty suspects more than doing nothing to prevent a distinctive suspect from standing out (Chapters 2 and 3). These data fit with a new model of eyewitness identification behaviour—the diagnostic-feature-detection model (Wixted & Mickes, 2014). The model suggests that people are better at discriminating between innocent and guilty suspects when they base their decisions on (diagnostic) facial features that differ between innocent and guilty suspects rather than on (non-diagnostic) facial features that innocent and guilty suspects share. According to this account, all three fair lineups encourage subjects to rely on diagnostic facial features, because the non-diagnostic distinctive feature either appears on every member (replication), or on none of the members (pixelation and block), and so subjects cannot use it to make an identification decision. In unfair lineups, by contrast, when the suspect is left to stand out, his distinctive feature evokes a strong memory signal—a feeling of familiarity. When people rely on the distinctive feature to make their identification decision, it impairs their ability to discriminate because both the innocent and guilty suspect share that feature. In short, research and theory suggest that when a culprit has a distinctive feature at the time of a crime, all three fair lineup procedures equally enhance eyewitness identification performance more than leaving the suspect to stand out.

The research in the preceding chapters tells us how fair lineup techniques affect identification behaviour when the culprit, who actually committed the crime, and the suspect, who may or may not have committed the crime, share the same distinctive feature. But what happens when the culprit does not have a distinctive feature, yet the police suspect does? Perhaps, for example, the police suspect is the culprit (i.e., he is guilty), but he has gained a new distinctive feature since committing the crime. Alternatively, perhaps the police suspect is not the culprit (i.e., he is innocent) and has a distinctive feature that the real culprit did not have. The police can never be sure if their suspect is guilty or innocent, but, in such cases, current police guidelines stipulate that the foils should still match the appearance of the suspect, for instance: "If the [witness'] description does not fit the suspect on some characteristic (e.g., the witness described dark hair, yet the suspect has light hair), then the fillers [i.e., foils] should match the suspect on that characteristic rather

than matching the description on that characteristic so that the suspect does not unduly stand out." (Technical Working Group for Eyewitness Evidence, 2003, p. 33) or "so far as possible, [the foils should] resemble the suspect in age, general appearance and position in life." (Police and Criminal Evidence Act 1984, Code D, 2011, p. 47). Although different guidelines suggest that suspects should not be left to stand out, there is some confusion about how best to achieve this goal. The Technical Working Group for Eyewitness Evidence (2003) in the US deems replication to be most appropriate, while the Police and Criminal Evidence Act 1984, Code D (2011) in England and Wales endorses concealment (i.e., pixelation or block techniques).

Although these guidelines exist, research is yet to examine how different lineup techniques for accommodating distinctive suspects affect eyewitness identification behaviour when the culprit does not have the distinctive feature during the crime. On the one hand, we might predict that replication, pixelation and block techniques will result in equivalent eyewitness identification accuracy when the culprit does not have the feature during the crime. According to the diagnostic-feature-detection model, all four lineups—replication, pixelation, block and do-nothing—should lead to similar levels of performance. The three fair lineups should have a similar cognitive effect because, regardless of whether the feature appears on every member (replication) or on none of the members (pixelation or block), subjects will discount the feature when making an identification decision. Similarly, yet somewhat unintuitively, the model also predicts that unfair lineups should not harm witnesses' ability to identify the culprit when the culprit does not have a distinctive feature during the crime. Of course, when the culprit doesn't have the feature, the witness doesn't have the opportunity to encode it. Thus, the feature on the face of the suspect in an unfair do-nothing lineup will not evoke a strong memory signal. And without a memory of the feature, subjects should give little weight to the non-diagnostic feature in the do-nothing lineups, so their ability to discriminate between innocent and guilty suspects should be similar on fair and unfair lineups. In short, the model predicts similar performance on all four lineups when the culprit does not have the feature during the crime.

On the other hand, we might predict that replication lineups will result in less accurate eyewitness identifications than both pixelation and block techniques when the culprit does not have the feature during the crime. In replication lineups, each

member of the lineup has a new distinctive feature. We know that recognition accuracy is harmed when face stimuli change between study and test, compared to when they remain the same (Ellis, 1975; Metzger, 2001; Patterson & Baddeley, 1977; see also Shapiro & Penrod, 1986 for a meta-analysis). Even the addition or removal of glasses (e.g., Terry, 1994, Experiment 1) and facial hair (e.g., Metzger, 1999; Terry, 1994, Experiment 1) can impair performance on face recognition and lineup tasks (Read, 1995). These findings fit with the encoding-specificity hypothesis, which suggests that performance on memory tasks is poorer when information available at encoding is different to the information available at retrieval (Tulving & Thomson, 1973). Together, this research suggests that eyewitness identification accuracy may be impaired when a previously unseen feature is replicated over the lineup members, compared to when that feature is concealed using pixelation or block techniques.

In this chapter, we examined how the lineup techniques for distinctive suspects influence eyewitness identification accuracy when the culprit does not have a distinctive feature during the crime, and compared this to performance on the same lineups when the culprit does have a feature during the crime. To this end, subjects watched a video of a mock crime being committed by a culprit without a distinctive feature, or by the same culprit with a distinctive feature. All subjects then attempted to identify the culprit from a replication, pixelation, block or do-nothing lineup and provided a confidence rating. We also collected subjects' descriptions of the culprit to explore the extent to which subjects freely recalled distinctive feature information, and to examine how subjects who failed to freely recall distinctive feature information performed on the four lineup types.

Method

Design

We used a 2 (culprit: non-distinctive, distinctive) × 4 (lineup type: replication, pixelation, block, do-nothing) × 2 (target: present, absent) mixed design, with target manipulated within subjects. The mixed design enabled us to collect two data points per subject: Each subject watched two mock crime videos and completed two lineup tasks (one target-present, one target-absent). Because we were interested in investigating general effects, we planned to collapse our data over the two crime

videos. We recruited as many subjects as possible before the end of summer term, aiming for at least 120 subjects with usable data in each of the eight between-subject conditions.

Subjects

We recruited 1,578 subjects via Amazon Mechanical Turk who each received $0.60. We excluded 115 people (7% in total; between 10–18 subjects in each of the eight between-subject conditions) who experienced technical difficulties while watching the video (n = 13, <1% in total), incorrectly answered an attention check question (n = 19, 1% in total), or stated that they had seen one of the videos before or had completed the study more than once (n = 83, 5% in total). The final sample size was 1,463, with 180–189 subjects in each of the between-subject cells (532 male, 866 female, 65 other or prefer not to say; age range = 16–77 years, M = 34.68, SD = 11.98). The majority of the sample self-identified as White (71.84%), the remainder identified as Asian (15.38%), Black (7.59%), Mixed (3.55%), or Other (1.09%), while 0.55% chose not to disclose their race or ethnicity.

Materials

Videos

We used four 30 s non-violent mock crime videos. Two of these videos were used in Chapters 2 and 3 and depicted different male culprits committing either a graffiti-attack or a mugging. In these videos, the *graffiti* culprit had a large bruise around his right eye and the *mugging* culprit had a large tribal tattoo on his right cheek. The other two videos were identical to the graffiti and mugging videos, but the culprit in each video did not have a distinctive feature. The distinctive and non-distinctive versions of each mock crime were filmed on the same day and were subsequently edited to eliminate any differences in timing that occurred during filming. This resulted in two sets of mock crime scenarios: graffiti (with black-eye, without black-eye) and mugging (with tattoo, without tattoo).

Lineups

We used the same lineup materials and construction strategy that we used in Chapters 2 and 3.

Procedure

Subjects were told that the study was about perception and memory and were randomly assigned into conditions, with the constraint that subject numbers were relatively equal in each condition. First, subjects watched a mock crime video (graffiti or mugging) in which the culprit either did or did not have a distinctive feature. The video was labelled as "Video A" and subjects were told to pay close attention because they would be asked questions about the content of the video later. Following this, we checked whether subjects had experienced any technical difficulties, such as problems playing the video or excessive buffering. Subjects were then given 2 min to describe the appearance of the male culprit in "Video A" in as much detail as possible. We told subjects: "It is important that you attempt to describe all of his different facial features," and stated that it was vital that they tried to describe the culprit for the full 2 min. Subjects were instructed to type everything that they could think of regarding the culprit's appearance into a box displayed in the centre of the screen. After 2 min, the study automatically advanced and the filler task began. The filler task consisted of 4 min of spatial reasoning questions. Subjects were then asked how confident they were that they would be able to recognise the culprit in "Video A", and rated their confidence on an 11-point Likert-type scale ranging from 0% (*completely uncertain*) to 100% (*completely certain*).

Next, subjects were told that they would be presented with a lineup, which may or may not contain the culprit. To prevent subjects who had watched the non-distinctive culprit from rejecting the lineup simply because the lineup members had distinctive features or concealed areas, we stated that the culprit's appearance may or may not have changed, and the lineup images may or may not have been digitally altered in some way. We told subjects that their task was to recognise the person they previously viewed in Video A. These instructions follow Technical Working Group for Eyewitness Evidence (1999) guidelines in the US which state that witnesses should be given an appearance-change warning, and also procedures in England and Wales whereby witnesses are told that the lineup images have been digitally modified (A. Monaghan, National VIPER User Group, personal communication, August 15, 2016; C. Wilkinson, Northamptonshire police, personal communication, August 17, 2016).

The next page displayed the lineup, either target-present or target-absent, composed of two rows of three photos. The lineup technique used (i.e., replication, pixelation, block or do-nothing), depended on the condition to which the subject had been randomly assigned. Subjects made an identification decision by clicking on the person who they believed was the culprit, or by choosing an option labelled "Not Present" if they thought the culprit was not in the lineup. Immediately after, subjects rated their confidence in their identification decision on an 11-point Likert-type scale ranging from 0% (*completely uncertain*) to 100% (*completely certain*). Subjects then answered a question assessing whether they had paid attention to the content of the video.

Next, subjects completed the same sequence of tasks again, this time viewing the alternative mock crime video (graffiti or mugging) and lineup format (target-present or target-absent). The video and tasks were labelled “Video B”. Subjects remained in the same distinctive condition to which they had been assigned such that subjects who had watched a distinctive culprit in “Video A” also watched a distinctive culprit in “Video B”, and vice versa. Subjects also remained in the same lineup technique condition to which they had been assigned. That is, subjects who had been presented with a replication lineup after “Video A”, for instance, were also presented with a replication lineup after “Video B”. The order of the videos and target conditions was counterbalanced. Finally, we asked subjects whether they believed they had seen either of the videos before or had completed the study more than once, and they answered several demographic questions.

Results

We conducted ROC analysis, examined subjects’ identification responses and analysed subjects’ descriptions of the culprit’s distinctive feature. We gathered further information by examining subjects’ ability to judge the accuracy of their suspect identification decisions.

ROC analysis

We constructed our ROC curves and calculated our *p*AUC statistics in the same way as in Chapters 2 and 3. In each set of ROC analysis, we set the specificity (1 – FAR) using the FAR range covered by the least extensive curve.

How do the lineup techniques for distinctive suspects influence eyewitness identification accuracy when the culprit does not have a distinctive feature during the crime? Figure 4.1a shows that when subjects watched a non-distinctive culprit, the ROC curves for the replication, pixelation, block and do-nothing lineups all lie on top of each other. This finding indicates that subjects' ability to discriminate between innocent and guilty suspects was similar on all four lineup types and fits with the prediction of the diagnostic-feature-detection model. The *p*AUCs (specificity = .968; see Table 4.1) did not differ significantly between replication and pixelation (D = 0.57, p = .57), replication and block (D = 0.32, p = .75), or pixelation and block (D = 0.16, p = .88) lineups. The *p*AUC for do-nothing lineups was also similar to the *p*AUC for replication (D = 0.29, p = .77), pixelation (D = 0.79, p = .43), and block (D = 0.54, p = .59) lineups.

In contrast, Figure 4.1b shows that when subjects watched a distinctive culprit, the ROC curves for the replication, pixelation, and block lineups lie on top of each other, while the curve for the do-nothing lineup lies below these, closer to the dashed chance line. This finding suggests that all three fair lineups were equally effective and all three fair techniques enhanced subjects' ability to discriminate between innocent and guilty suspects more than doing nothing to prevent the distinctive suspect from standing out. This pattern of results replicates previous findings (e.g., Chapters 2 and 3) and fits with the diagnostic-feature-detection model. The *p*AUCs (specificity = .917) did not differ significantly between replication (*p*AUC = 0.030, 95% CI: 0.019, 0.041) and pixelation (*p*AUC = 0.025, 95% CI: 0.013, 0.037, D = 0.61, p = .54), replication and block (*p*AUC = 0.027, 95% CI: 0.018, 0.038, D = 0.35, p = .72), or pixelation and block (D = 0.30, p = .76) lineups. But the *p*AUC for do-nothing lineups (*p*AUC = 0.009, 95% CI: 0.005, 0.014) was significantly smaller than the *p*AUC for replication (D = 3.53, p < .001), pixelation (D = 2.37, p = .02) and block (D = 3.28, p = .001) lineups. Figure 4.1b also shows that subjects' willingness to identify the suspect was increased in the do-nothing lineups, compared to the replication, pixelation and block lineups, because the do-nothing curve extends further right than the curves for the three fair lineups, which reflects a larger hit rate and false alarm rate in the do-nothing lineup condition.

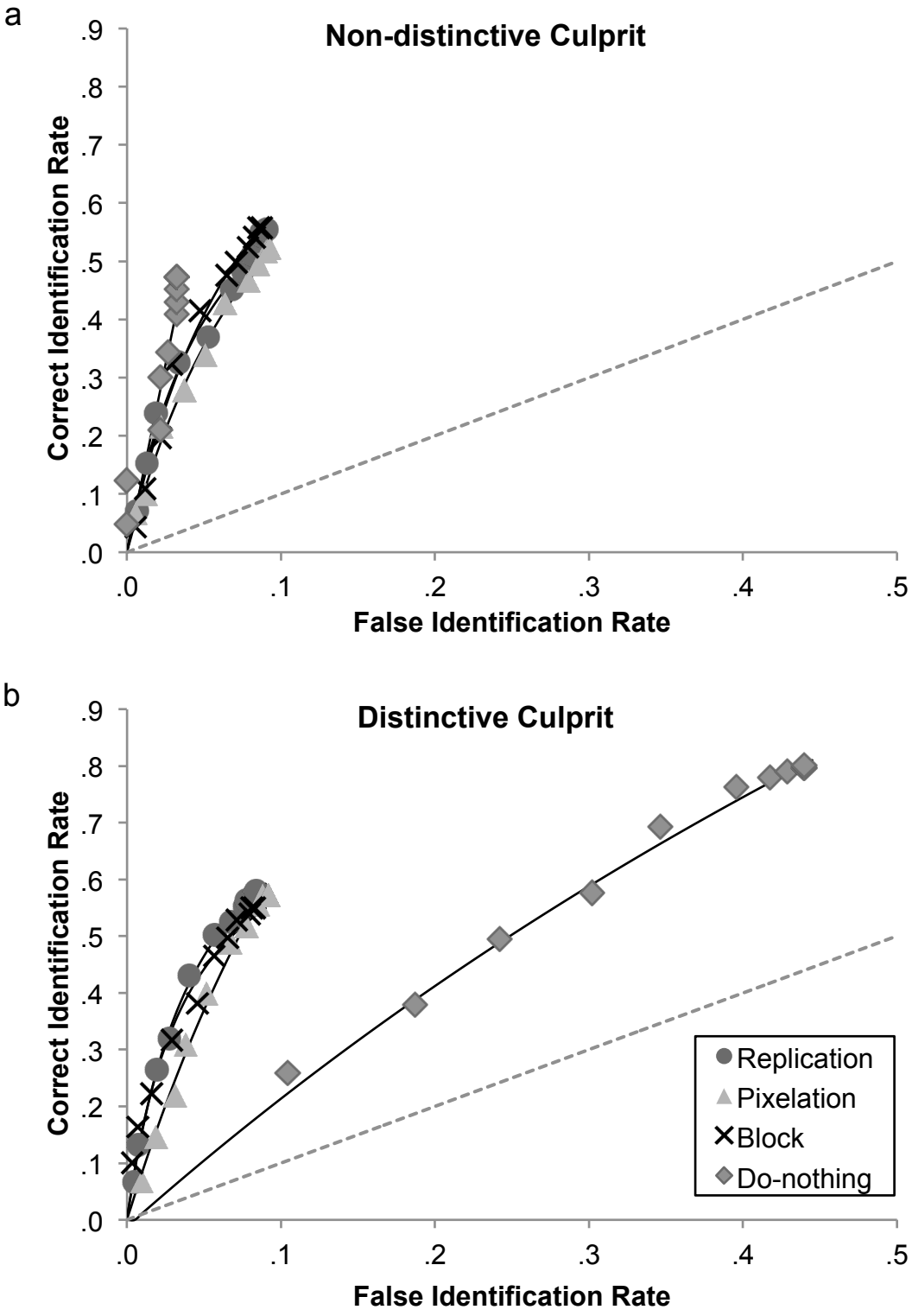

Figure 4.1. Receiver operating characteristic (ROC) curves for the replication, pixelation, block, and do-nothing lineups in subjects who had watched (a) a non-distinctive or (b) a distinctive culprit. The dashed lines represent chance-level performance.

Table 4.1
Partial Area Under the Curve (pAUC) Statistics [and 95% Confidence Intervals]

Lineup type	Non-distinctive culprit	Distinctive culprit
Replication	0.006 [0.003, 0.011]	0.006 [0.002, 0.012]
Pixelation	0.005 [0.002, 0.009]	0.003 [0.001, 0.009]
Block	0.005 [0.002, 0.012]	0.006 [0.003, 0.010]
Do-nothing	0.007 [0.004, 0.014]*	0.001 [0.001, 0.002]*

Note. Specificity (1 – FAR) = .968, which was set using the FAR range of the least extensive curve.
* *p*AUCs differ at $p = .02$.

The finding that all four lineups—fair and unfair—led to similar performance when subjects viewed the non-distinctive culprit is predicted by the diagnostic-feature-detection model. Yet we do not know if this was because all three fair lineup techniques were harmful to performance, or if this was because unfair lineups were not harmful to performance. To check this, we compared the identification performance of subjects who had watched the non-distinctive culprit with subjects who had watched the distinctive culprit on each lineup type. Figure 4.2 shows that subjects performed similarly on the three fair lineups, regardless of whether the culprit had the feature during the crime; it was only performance on the unfair do-nothing lineups that differed. Specifically, ability to discriminate between innocent and guilty suspects was better on do-nothing lineups when the culprit did not have the feature during the crime, compared to when the culprit did have the feature during the crime. The *p*AUCs (specificity = .968; see Table 4.1) did not differ significantly between subjects who had watched a non-distinctive or a distinctive culprit on the replication ($D = 0.14$, $p = .89$), pixelation ($D = 0.51$, $p = .61$), or block ($D = 0.17$, $p = .87$) lineups. Whereas, the *p*AUC for subjects who had watched a non-distinctive culprit was significantly larger than the *p*AUC for subjects who watched a distinctive culprit on the do-nothing lineups ($D = 2.32$, $p = .02$). This illustrates that adding a replicated feature, an area of pixelation, or a black block to the lineup faces had no depreciable effect on identification accuracy. Therefore the equivalent performance on the fair and unfair lineups after subjects had watched the non-distinctive culprit cannot be attributed to poor performance on the three fair lineup techniques. Instead, ability to discriminate between innocent and guilty suspects on unfair lineups was better when subjects did not have the opportunity to encode the

feature during the crime, compared to when subjects had a memory of that feature. Or, to put it another way, ability to discriminate between innocent and guilty suspects on unfair lineups was only impaired when subjects had the opportunity to encode the distinctive feature during the crime.

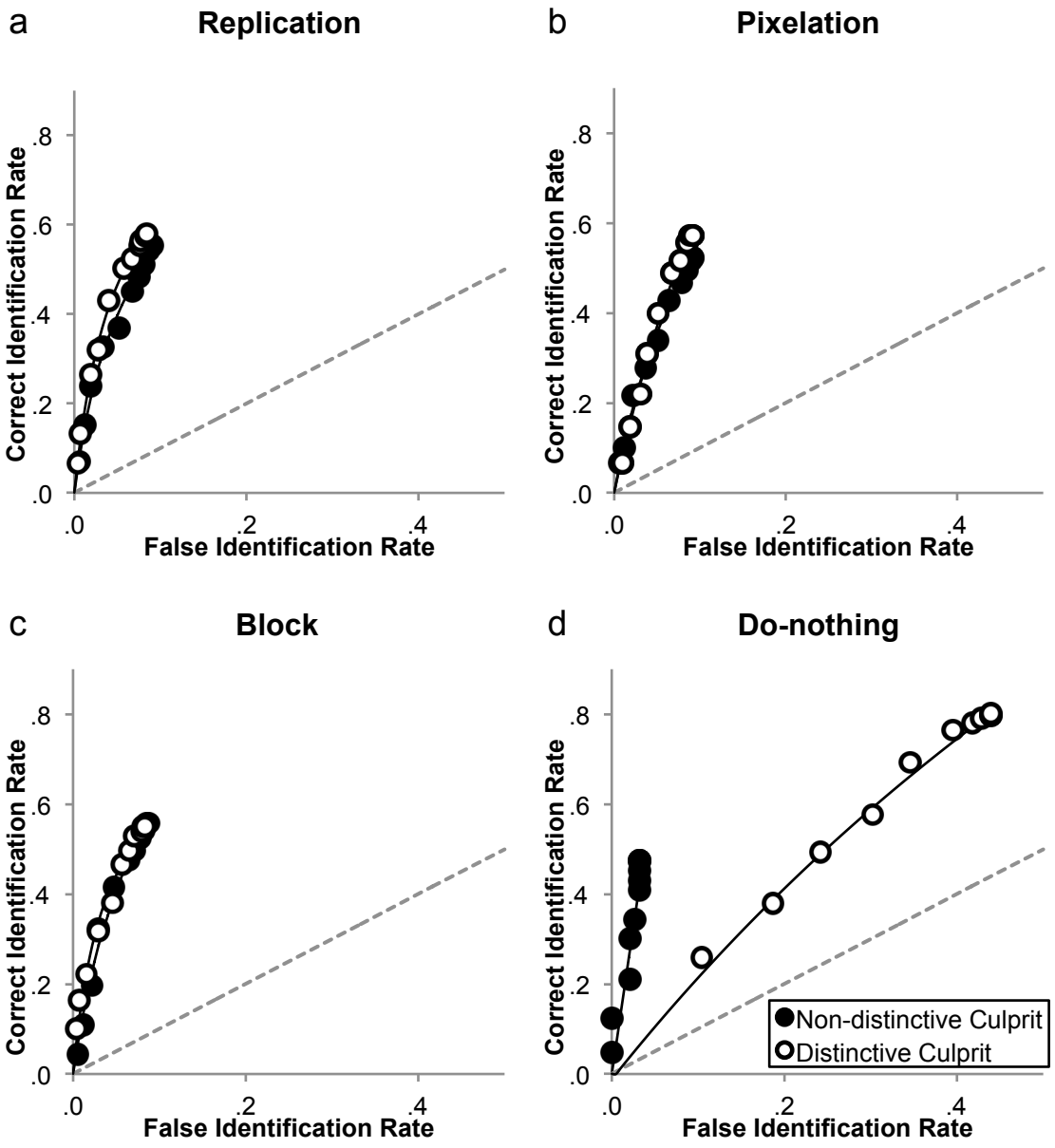

Figure 4.2. Receiver operating characteristic (ROC) curves for the (a) replication, (b) pixelation, (c) block, and (d) do-nothing lineups in subjects who had watched a non-distinctive or a distinctive culprit. The dashed lines represent chance-level performance.

We also fit a signal detection process model to our data (Wixted & Mickes, 2014; see Chapter 1 for a description of the model). The results of our atheoretical

*p*AUC analysis corresponded to the results obtained by fitting the theoretical model to the data (cf. Lampinen, 2016; see Appendix G).

Identification responses

To further understand how the lineup techniques for distinctive suspects influence eyewitness identification performance we calculated the proportion of suspect identifications, foil identifications and lineup rejections (i.e., "Not Present" responses) in each lineup type. Figure 4.3 shows the identification responses made in replication, pixelation, block and do-nothing (a) target-present and (b) target-absent lineups, by subjects who had watched a non-distinctive or distinctive culprit. For target-absent lineups, we calculated the number of innocent suspect and foil identifications in the same way as in Chapters 2 and 3.

Target-present lineups. Figure 4.3a shows that subjects who had watched a non-distinctive culprit performed similarly on all four lineups, whereas subjects who had watched a distinctive culprit performed similarly on the three fair lineups, but made more guilty suspect identifications on the unfair do-nothing lineups. A 2 (culprit: non-distinctive, distinctive) × 4 (lineup type: replication, pixelation, block, do-nothing) × 3 (identification response: guilty suspect, foil, incorrect rejection) hierarchical loglinear analysis revealed a significant three-way interaction, indicating that the appearance of the culprit and lineup type influenced identification responses, χ^2 (6, N = 1,463) = 33.07, $p < .001$. Two 4 (lineup type: replication, pixelation, block, do-nothing) × 3 (identification response: guilty suspect, foil, incorrect rejection) two-way chi-square analyses indicated that when subjects watched a non-distinctive culprit, lineup technique did not influence their identification responses, χ^2 (6, N = 733) = 7.40, $p > .250$, Cramer's V = .07. Conversely, when subjects watched a distinctive culprit, lineup technique influenced their identification responses, χ^2 (6, N = 730) = 32.92, $p < .001$, Cramer's V = .15. When subjects watched a distinctive culprit, they responded similarly on the three fair lineup types, but they made more guilty suspect IDs (z = 3.00, $p < .01$), but fewer foil IDs (z = –2.86, $p < .01$) and fewer rejections (z = –2.63, $p < .01$) than expected on the do-nothing lineups. Specifically, three 2 (lineup type) × 2 (identification response: guilty suspect, foil) two-way chi-square analyses indicated that when subjects made a selection, they were over 3 times more likely to identify the guilty suspect from the unfair do-nothing lineups than the fair lineups, replication and do-nothing, χ^2 (1, N =

305) = 14.67, $p < .001$, OR = 3.16, 95% CI [1.67, 6.18]; pixelation and do-nothing, χ^2 (1, $N = 310$) = 13.89, $p < .001$, OR = 3.09, 95% CI [1.62, 6.08]; block and do-nothing, χ^2 (1, $N = 310$) =15.66, $p < .001$, OR = 3.26, 95% CI [1.73, 6.38].

Target-absent lineups. Figure 4.3b also shows that subjects who had watched the non-distinctive culprit performed similarly on all four lineups, whereas subjects who had watched the distinctive culprit performed similarly on the three fair lineups, but made more innocent suspect identifications on the unfair do-nothing lineups. A 2 (culprit: non-distinctive, distinctive) × 4 (lineup type: replication, pixelation, block, do-nothing) × 3 (identification response: innocent suspect, foil, correct rejection) hierarchical loglinear analysis revealed a significant three-way interaction, indicating that the appearance of the culprit and lineup type influenced identification responses, χ^2 (6, $N = 1{,}463$) = 62.97, $p < .001$. Two 4 (lineup type: replication, pixelation, block, do-nothing) × 3 (identification response: innocent suspect, foil, correct rejection) two-way chi-square analyses indicated that when subjects watched a non-distinctive culprit, lineup technique did not influence their identification responses, χ^2 (6, $N = 733$) = 7.68, $p > .250$, Cramer's $V = .07$. Conversely, when subjects watched a distinctive culprit, lineup technique influenced their identification responses, χ^2 (6, $N = 730$) = 120.88, $p < .001$, Cramer's $V = .29$. When subjects watched a distinctive culprit, they responded similarly on the three fair lineup types and made fewer innocent suspect IDs than expected (replication: $z = –2.91$, $p < .01$, pixelation: $z = –2.64$, $p < .01$, block: $z = –3.01$, $p < .01$), but they made more innocent suspect IDs ($z = 8.58$, $p < .001$), fewer foil IDs ($z = –3.28$, $p < .01$) and fewer rejections ($z = –2.34$, $p < .05$) than expected on the do-nothing lineups. Specifically, three 2 (lineup type) × 2 (ID response: innocent suspect, foil) two-way chi-square analyses indicated that when subjects made a selection, they were over 9 times more likely to identify the innocent suspect from the unfair do-nothing lineups than the fair lineups, replication and do-nothing, χ^2 (1, $N = 213$) = 50.43, $p < .001$, OR = 9.54, 95% CI [4.75, 20.16]; pixelation and do-nothing, χ^2 (1, $N = 220$) = 52.82, $p < .001$, OR = 9.64, 95% CI [4.88, 20.00]; block and do-nothing, χ^2 (1, $N = 216$) = 51.46, $p < .001$, OR = 9.17, 95% CI [4.63, 19.07].

In sum, these results align with the findings of the ROC analysis and suggest all four lineup techniques—replication, pixelation, block, and do-nothing—result in a similar pattern of identification responses when the culprit does not have a

distinctive feature during the crime. The pattern of identification responses made on the fair lineups was similar in subjects who had watched the non-distinctive and distinctive culprit, which illustrates that adding something new (i.e., a replicated feature, an area of pixelation, or a black block) to the lineup faces had no depreciable effect on performance. Finally, subjects were only prone to picking the distinctive suspect on the unfair lineups when they had the opportunity to encode the distinctive feature during the crime.

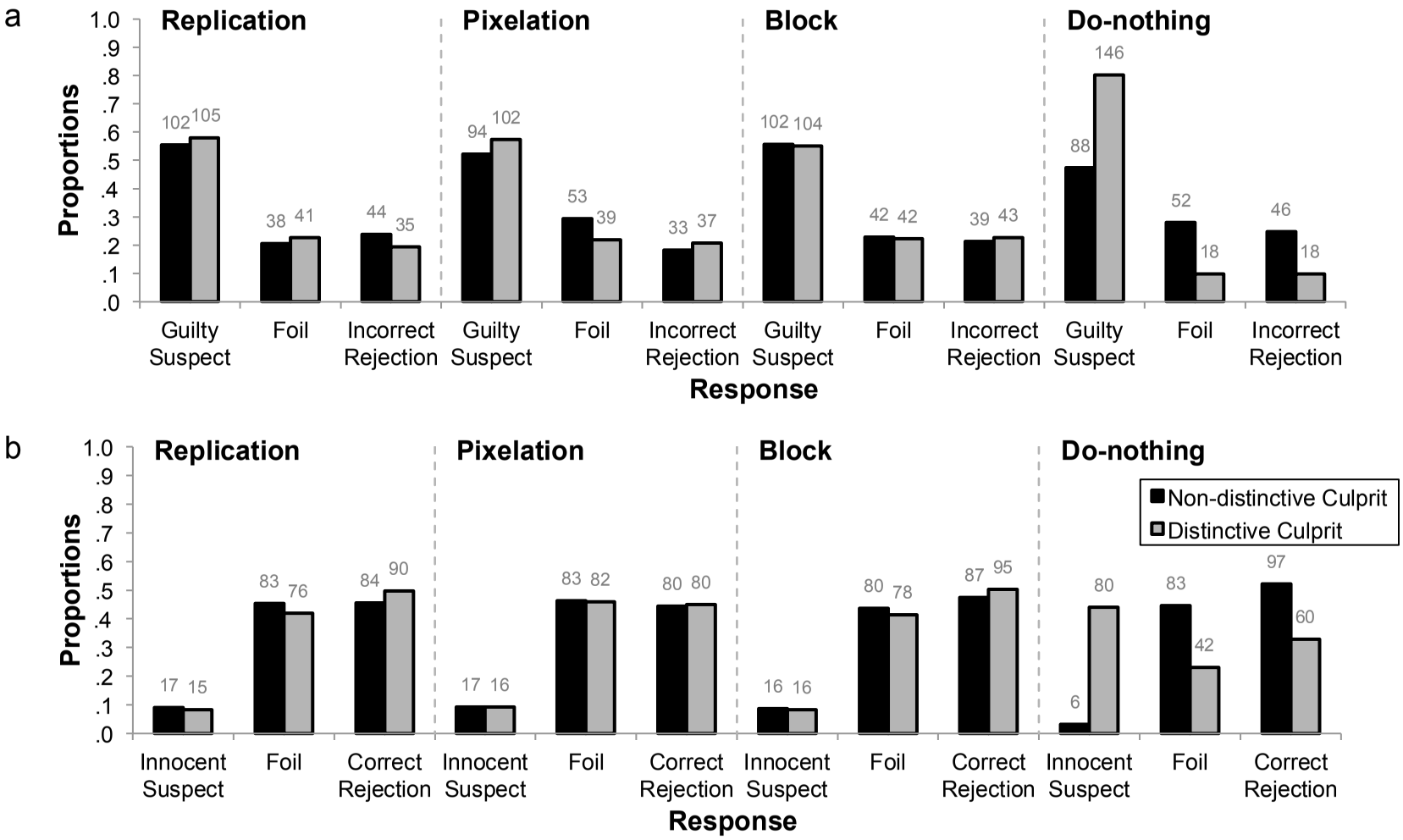

Figure 4.3. Identification responses made in replication, pixelation, block, and do-nothing (a) target-present and (b) target-absent lineups, by subjects who had watched a non-distinctive or a distinctive culprit. Data labels are absolute frequencies.

Descriptions

So far our analyses have shown that fair (replication, pixelation and block) and unfair (do-nothing) lineups resulted in equivalent eyewitness accuracy when the suspect had a new distinctive feature that was not present during the crime. But police officers can never be certain that a suspect's distinctive feature is new, because they are reliant on the witness's description of the culprit. It is possible that a witness could fail to report a culprit's distinctive feature that was present during the crime. A witness, for instance, might have encoded the feature without conscious awareness (e.g., Lewicki, Hill, & Czyzewska, 1992), or might be unable to freely recall information about the feature because it is difficult to verbally articulate what has been learnt (e.g., Lewicki, Czyzewska, & Hoffman, 1987; R. C. L. Lindsay, Martin, & Webber, 1994; Meissner, Sporer, & Schooler, 2007; Nelson, 1978). If the culprit had the feature at the time of the crime, but the witness failed to report it, then the witness may still have encoded the feature and, consequently, may perform differently on the lineups for distinctive suspects compared to a witness who did not have the opportunity to encode the feature at all (i.e., subjects who watched our non-distinctive videos).

To this end, we analysed subjects' descriptions of the culprit to examine (a) the extent to which subjects who had watched a distinctive culprit freely recalled distinctive feature information, and (b) how subjects who had watched a distinctive culprit but failed to freely recall distinctive feature information performed on the four lineup types. This information can provide us with a more comprehensive understanding of the most appropriate lineup technique(s) to use when a witness does not describe a distinctive feature, but the suspect has a feature at the time of the lineup.

We devised a coding scheme to assess the type and level of detail subjects freely recalled about the culprit's distinctive feature. Four coders, who were blind to purpose of the study and the condition to which subjects had been assigned, completed the coding independently. To assess interrater reliability, we randomly selected 5% of the descriptions to be coded by all four coders and computed Siegel and Castellan's kappa for each coder pair. The average kappa indicated substantial agreement, $\kappa = 0.80$ (Landis & Koch, 1977). All coding discrepancies were resolved

through discussion between the first author and the four coders. The four coders then coded 25% of the remaining descriptions, each.

Subjects' reports of the distinctive feature

Table 4.2 shows the frequency of descriptions in each coding category for non-distinctive and distinctive culprits. Subjects clearly engaged with the task: only 121 descriptions (4.14% of the total) were missing or completed incorrectly (codes 6 and 7) and the average word count of the correctly completed descriptions (codes 0–5) was 30.89 (*SD* = 16.29) and 33.73 (*SD* = 15.87) for the non-distinctive and the distinctive culprits, respectively. Table 4.2 shows that the vast majority of subjects who watched a distinctive culprit freely recalled some information about the distinctive feature. Almost half of the distinctive descriptions included specific details about the feature, such as location or shape (codes 3–5). Conversely, fewer than 10% of the distinctive descriptions contained information about the culprit's general appearance but failed to report the distinctive feature (code 0). Although this category represents a small minority of all subjects, this illustrates that it is possible for people to fail describe a prominent distinctive facial feature, even when they have been prompted to describe it and it was viewed under relatively good encoding conditions.

Table 4.2
Percentages (and Frequencies) of Descriptions in Each Coding Category

Code	Non-distinctive culprit		Distinctive culprit	
0 = did not describe the feature	95.16	(1,395)	9.45	(138)
1 = described something to do with the feature	0.00	(0)	17.53	(256)
2 = described the feature correctly	0.00	(0)	22.05	(322)
3 = described the feature correctly *with specific location*	0.00	(0)	36.30	(530)
4 = described the feature correctly *in detail*	0.00	(0)	2.81	(41)
5 = described the feature correctly *with specific location and in detail*	0.00	(0)	8.42	(123)
6 = completed the task incorrectly	4.16	(61)	2.67	(39)
7 = did not write anything	0.68	(10)	0.75	(11)
Total	100.00	(1,466)	100.00	(1,460)

Performance by subjects who failed to recall the distinctive feature

We examined identification responses on the four lineup types in (a) subjects who didn't describe the feature because they had watched a non-distinctive culprit (non-distinctive culprit, code 0), (b) subjects who failed to describe the feature even though they had watched a distinctive culprit (distinctive culprit, code 0), and (c) subjects who described the feature after watching a distinctive culprit (distinctive culprit, codes 1–5). Figure 4.4 shows the identification responses made in replication, pixelation, block and do-nothing (a) target-present and (b) target-absent lineups, by subjects' descriptions.[2] Again, for target-absent lineups, we calculated the number of innocent suspect and foil identifications in the same way as in Chapters 2 and 3.

Target-present lineups. Figure 4.4a shows that the pattern of identification responses to the lineup types differed depending on the subject's description. A 3 (feature: non-distinctive, not described, described) × 4 (lineup type: replication, pixelation, block, do-nothing) × 3 (identification response: guilty suspect, foil, incorrect rejection) hierarchical loglinear analysis revealed a significant three-way interaction, χ^2 (12, N = 1,402) = 41.57, $p < .001$. To examine this further we conducted four, 3 (feature: non-distinctive, not described, described) × 3 (identification response: guilty suspect, foil, incorrect rejection) two-way chi-square analyses, one for each lineup type.[3] Those who failed to describe the feature tended to make fewer guilty suspect IDs ($z = –2.17$, $p < .05$) and more rejections ($z = 2.43$, $p < .05$) than those who described the feature or those who had watched the non-distinctive culprit, but this was only statistically significant in replication lineups, χ^2 (4, N = 349) = 15.59, $p = .004$, Cramer's $V = .15$. Generally speaking, regardless of the content of their description, subjects performed similarly on the fair lineup types, pixelation: χ^2 (4, N = 345) = 7.18, $p = .13$, Fisher's exact test $p = .10$, Cramer's V = .10; block: χ^2 (4, N = 358) = 1.79, $p > .250$, Cramer's $V = .05$. Subjects did, however, make a different pattern of identification responses on the do-nothing lineup depending on the content of their description, χ^2 (4, N = 350) = 42.95, $p < .001$,

[2] Because only a minority of subjects who had watched a distinctive culprit failed to freely recall information about the distinctive feature, there were too few observations to conduct ROC analysis or fit a signal detection process model to these data.

[3] Fisher's exact test is reported for tests in which more than 20% of cells had expected frequencies of less than 5 (Field, Miles, & Field, 2012).

Fisher's exact test $p < .001$, Cramer's $V = .25$. Three 2 (feature) × 2 (identification response: guilty suspect, foil) two-way chi-square analyses indicated that when subjects made an identification, subjects who described the feature were not significantly more likely to identify the guilty suspect compared to those who failed to describe the feature, χ^2 (1, $N = 155$) = 2.43, $p = .12$, Fisher's exact test $p = .14$, OR = 2.90, 95% CI [0.46, 13.29], but they were 5.91 times more likely to identify the guilty suspect compared to those who watched the non-distinctive culprit, χ^2 (1, $N = 273$) = 29.54, $p < .001$, OR = 5.91, 95% CI [2.88, 12.97]. Subjects who failed to describe the feature were not significantly more likely to identify the guilty suspect compared to those who watched the non-distinctive culprit, χ^2 (1, $N = 146$) = 1.13, $p > .250$, Fisher's exact test $p > .250$, OR = 2.02, 95% CI [0.50, 11.82]. This highlights that the proportion of subjects who failed to describe the feature but identified the distinctive suspect in the do-nothing lineup was midway between the proportion of subjects who described the feature, and those who did not have the opportunity to encode the feature at all. Put simply, some subjects who watched a distinctive culprit but failed to describe the feature were still influenced by the distinctive feature in the do-nothing lineups.

Target-absent lineups. Figure 4.4b shows that subjects performed similarly on the three fair lineups, but some subjects who failed to describe the feature were influenced by the distinctive feature in the do-nothing lineups. A 3 (feature: non-distinctive, not described, described) × 4 (lineup type: replication, pixelation, block, do-nothing) × 3 (identification response: innocent suspect, foil, correct rejection) hierarchical loglinear analysis revealed a significant three-way interaction, χ^2 (12, $N = 1{,}403$) = 68.78, $p < .001$.[4] To examine this further we conducted four 3 (feature: non-distinctive, not described, described) × 3 (identification response: innocent suspect, foil, correct rejection) two-way chi-square analyses, one for each lineup type. Regardless of the content of their description, subjects performed similarly on each of the three fair lineup types, replication: χ^2 (4, $N = 346$) = 0.62, $p > .250$, Cramer's $V = .03$; pixelation: χ^2 (4, $N = 346$) = 0.17, $p > .250$, Cramer's $V = .02$;

[4] Because some cells had frequencies that were not greater than 1, we also conducted the same hierarchical loglinear analysis, but we collapsed the data over the three fair lineup techniques. The results were the same regardless of whether we collapsed the data over the fair lineup techniques or not.

block: χ^2 (4, N = 359) = 1.36, p > .250, Cramer's V = .04. Subjects, however, made a different pattern of identification responses on the do-nothing lineup depending on the content of their description, χ^2 (4, N = 352) = 95.73, p < .001, Cramer's V = .37. Three 2 (feature) × 2 (identification response: innocent suspect, foil) two-way chi-square analyses indicated that when subjects made an identification, subjects who described the feature were 6.61 times more likely to identify the innocent suspect compared to those who failed to describe the feature, χ^2 (1, N = 115) = 6.48, p = .007, Fisher's exact test p = .02, OR = 6.61, 95% CI [1.11, 70.40] and 35.05 times more likely to identify the innocent suspect compared to those who watched the non-distinctive culprit, χ^2 (1, N = 192) = 78.32, p < .001, OR = 35.05, 95% CI [12.79, 121.44]. Subjects who failed to describe the feature were also 5.17 times more likely to identify the innocent suspect compared to those who watched the non-distinctive culprit, χ^2 (1, N = 93) = 3.84, p = .05, Fisher's exact test p = .11, OR = 5.17, 95% CI [0.41, 41.61]. Again, this highlights that some of the subjects who watched a distinctive culprit but failed to describe the feature were still influenced by the distinctive feature in the do-nothing lineups.

Overall, the findings presented here provide preliminary evidence that using unfair lineups may be a risky strategy. Some of the subjects who watched a distinctive culprit but failed to freely recall the feature still responded as if they recognised the feature because they were prone to identifying the distinctive suspect on the unfair lineups. Generally speaking, the three fair lineups led to a similar pattern of identification responses, regardless of the description provided by the subject. In the fair target-present lineups, however, subjects who failed to describe the distinctive feature tended to be less likely to correctly identify the guilty suspect. Perhaps these subjects attended to the culprit's face for a shorter amount of time during the video, or perhaps they were not fully engaged with the task; they didn't make the effort to write a complete description or take care on the lineup. Whatever the reason for the overall poorer accuracy in target-present lineups, the key point is that all three fair lineup techniques—replication, pixelation and block—seem to be equally effective at promoting accurate eyewitness identifications even when the witness fails to report a culprit's distinctive feature that was present during the crime.

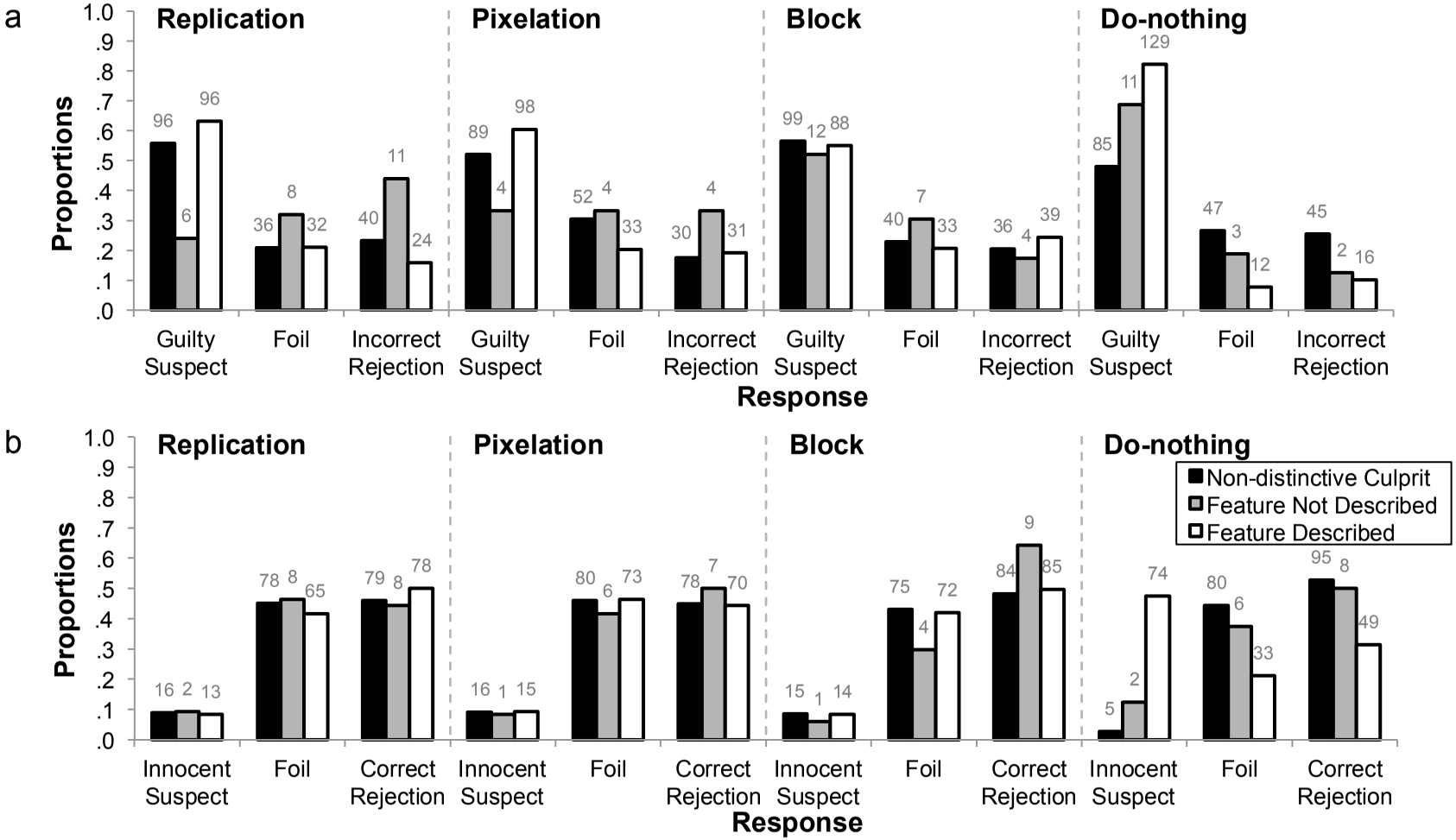

Figure 4.4. Identification responses made in replication, pixelation, block, and do-nothing (a) target-present and (b) target-absent lineups, by subjects who had watched a non-distinctive culprit, watched a distinctive culprit but failed to describe the feature, or watched a distinctive culprit and described the feature. Data labels are absolute frequencies.

Confidence and accuracy

Our analyses so far provide useful information for policymakers (e.g., legislators and police chiefs) who decide what type of lineup should be used. When a case gets to court, however, information that is useful for legal decision makers (e.g., judges and jurors) is whether witnesses express low confidence in identifications that are unlikely to be accurate and higher levels of confidence in identifications that are likely to be accurate (Mickes, 2015). Previous research shows that when the culprit has a feature during the crime, suspect identifications are significantly less accurate at every level of confidence on unfair do-nothing lineups compared to the three fair lineups (Chapters 2 and 3). According to the diagnostic-feature-detection account, this is because it is not clear that the distinctive feature is unhelpful in do-nothing lineups, so subjects fail to lower their confidence judgement despite using the (non-diagnostic) feature to make their identification decision. The fair lineups prevent subjects from relying on the non-diagnostic feature and so subjects are able to assess the likely accuracy of their memories and only make decisions with high confidence when they are likely to be correct. What's still unknown, however, is whether people are able to assess the likely accuracy of their memories on the lineups for distinctive suspects when the culprit does not have the feature during the crime, and if they can do this to the same extent as people who witnessed the crime committed by a culprit with a distinctive feature.

To test this, we constructed confidence-accuracy curves in the same way as in Chapters 2 and 3. The frequencies of identification responses in each confidence bin are presented in Appendix H. How do the lineup techniques for distinctive suspects influence suspect identification accuracy at each level of confidence when the culprit does not have a distinctive feature during the crime? Figure 4.5 shows the confidence-accuracy curves for each lineup type in subjects who watched (a) a non-distinctive or (b) a distinctive culprit. When the error bars do not overlap, this illustrates reliable differences between the lineup techniques (Sauer et al., 2010). When the culprit did not have the feature during the crime (Figure 4.5a) suspect identifications were perfectly accurate at almost every level of confidence on the unfair do-nothing lineups. At the higher levels of confidence (i.e., 50–60%, 70–80% and 90–100% certain) all four lineups resulted in a similar proportion of correct suspect identifications. Conversely, when the culprit did have the feature during the

crime (Figure 4.5b), suspect identifications were significantly less accurate at almost every level of confidence on the unfair do-nothing lineups compared to the three fair lineups, while all three fair lineups resulted in a similar proportion of correct suspect identifications.

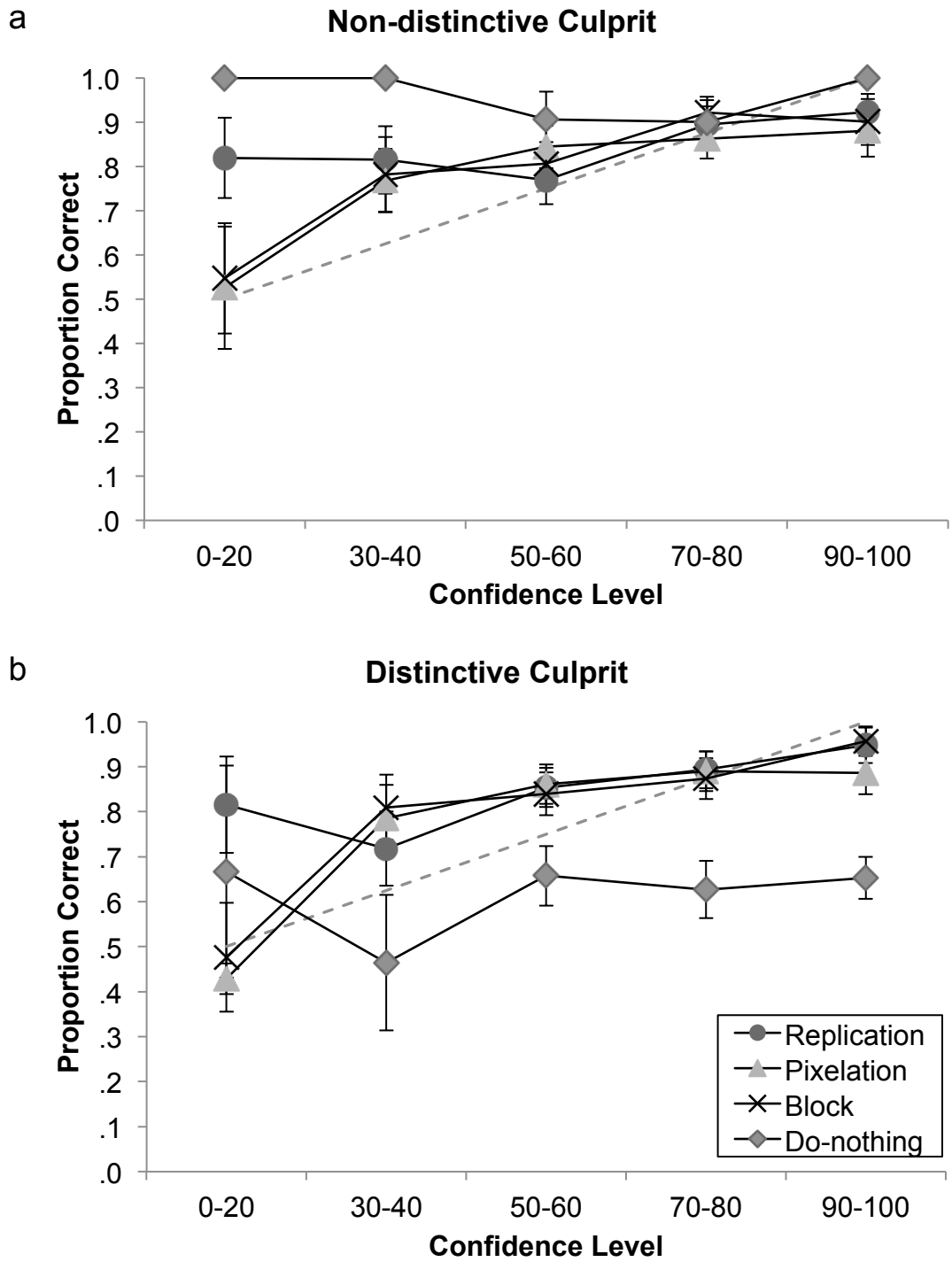

Figure 4.5. Confidence-accuracy curves for suspect identifications in the fair (replication, pixelation, block) and unfair (do-nothing) lineups by subjects who had watched (a) a non-distinctive or (b) a distinctive culprit. Error bars indicate ±1 SE. The dashed lines represent chance accuracy at the lowest confidence bin (i.e., 0–20) and perfect accuracy at the highest confidence bin (i.e., 90–100).

How does eyewitness identification accuracy at each level of confidence differ on lineups for distinctive suspects when the culprit does not have the feature during the crime, compared to when the culprit does have the feature during the crime? Figure 4.6 shows that accuracy at every level of confidence on the (a) replication, (b) pixelation, and (c) block lineups was the same, regardless of whether the culprit had the distinctive feature during the crime. This illustrates that adding a replicated feature, an area of pixelation, or a black block to the lineup faces had no depreciable effect on subjects' ability to judge the likely accuracy of their identification decision. Conversely, suspect identifications on the unfair do-nothing lineups (Figure 4.6d) were significantly more accurate at every level of confidence when the culprit did not have the feature during the crime compared to when the culprit did have the feature during the crime. This suggests that accuracy at each level of confidence is only impaired on unfair lineups when subjects had the opportunity to encode the distinctive feature during the crime.

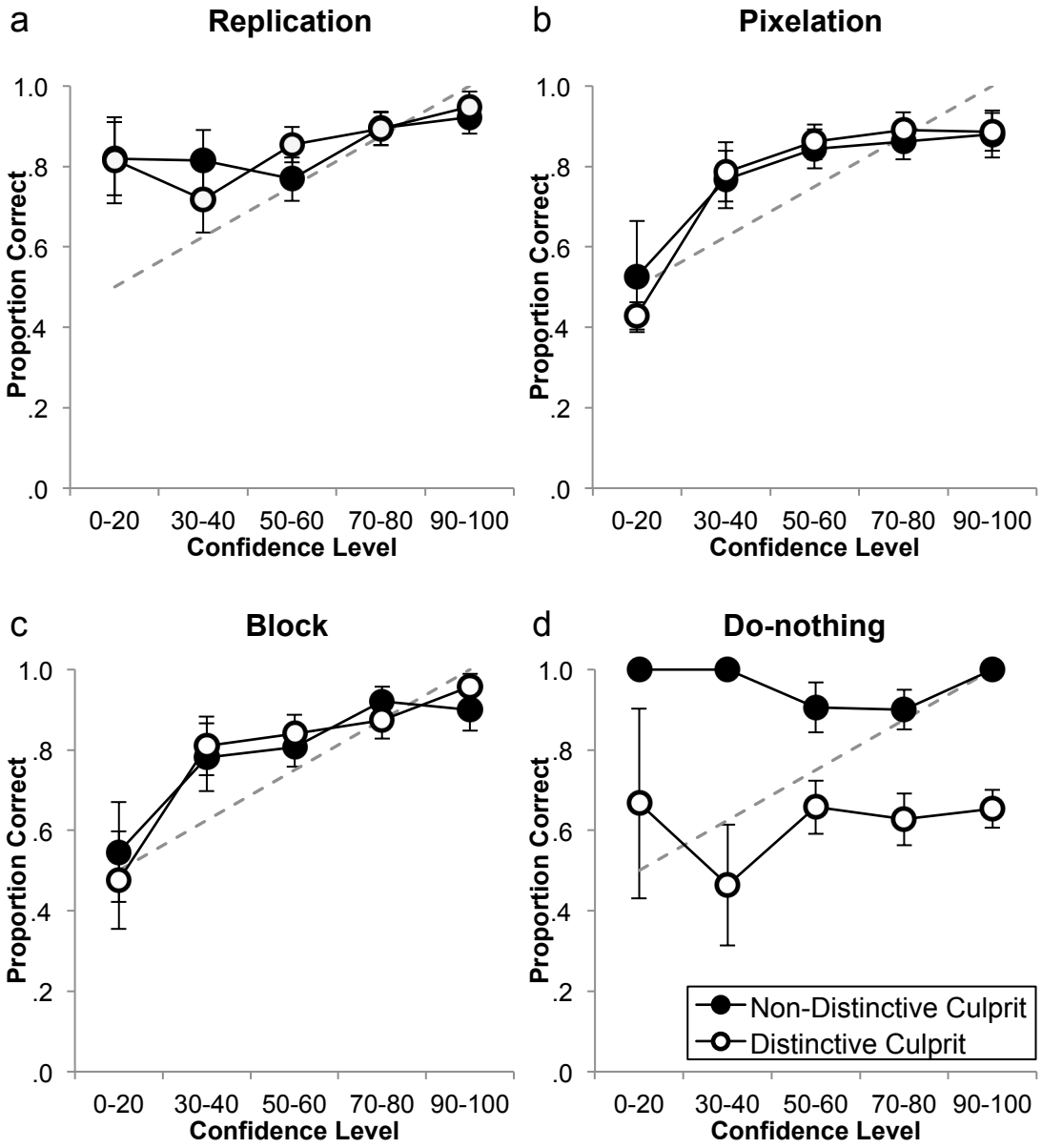

Figure 4.6. Confidence-accuracy curves for suspect identifications in the (a) replication, (b) pixelation, (c) block, and (d) do-nothing lineups by subjects who had watched a non-distinctive or a distinctive culprit. Error bars indicate ±1 SE. The dashed lines represent chance accuracy at the lowest confidence bin (i.e., 0–20) and perfect accuracy at the highest confidence bin (i.e., 90–100).

Discussion

We examined how the lineup techniques for distinctive suspects influence eyewitness identification accuracy when the culprit does not have a distinctive feature during the crime, and compared this to performance on the same lineups when the culprit does have a feature during the crime. When the culprit did not have the feature during the crime, all four lineup techniques (replication, pixelation, block

and do-nothing) were equally effective at promoting accurate eyewitness identifications and ensuring that identifications made with high-confidence were likely to be highly accurate. Performance on the fair (replication, pixelation, and block) lineups was the same, regardless of whether the culprit had the distinctive feature during the crime, but accuracy was better on unfair (do-nothing) lineups when the culprit did not have the feature during the crime compared to when the culprit did have the feature during the crime.

At first glance, our results seem to contradict a vast body of psychological literature and current police guidelines that warn against the dangers of leaving suspects to stand out (e.g., Fitzgerald et al., 2013; Police and Criminal Evidence Act 1984, Code D, 2011; Technical Working Group for Eyewitness Evidence, 1999; Wells, Leippe, & Ostrom, 1979). Indeed, our data suggest that simply leaving the suspect to stand out is not an inherently dangerous strategy: subjects' ability to discriminate between innocent and guilty suspects was similar in fair and unfair lineups when subjects had watched a non-distinctive culprit. According to the diagnostic-feature-detection account, this is because subjects who watched a non-distinctive culprit did not use the (non-diagnostic) distinctive feature to make an identification decision on any of the four lineup types (Wixted & Mickes, 2014). In the fair lineups, the feature either appeared on every member (replication) or on none of the members (pixelation or block) so could not be used in the identification decision, and in the unfair lineups, the feature did not evoke a strong memory signal because subjects had not encoded it. We found that it was only when subjects had watched a distinctive culprit that unfair lineups—compared to fair lineup—impaired identification accuracy. In this case, the suspect's distinctive feature evoked a strong memory signal and subjects used the feature to make their identification decision. This impaired identification accuracy because the feature was non-diagnostic; it was something that both the innocent and guilty suspect shared. In short, our research indicates that it is leaving the suspect to stand out in a way that is *consistent with the witness's memory of the culprit* that impairs identification accuracy and distorts confidence judgements.

Should we recommend, then, that leaving a distinctive suspect to stand out is acceptable in circumstances when a witness does not mention a distinctive feature in their description of the culprit? Our analysis of subjects' descriptions suggests no.

After watching a distinctive culprit, some subjects—albeit a minority—provided descriptions of the culprit's general appearance but failed to report his prominent distinctive feature. Thus, there could be multiple reasons why a real eyewitness's description does not include a distinctive feature. Yes, it is possible that the witness simply did not have the opportunity to encode the feature because the culprit did not have a feature during the crime (akin to our non-distinctive culprit condition). Equally, however, the culprit may have had the feature during the crime and the witness may have encoded the feature without conscious awareness (Lewicki et al., 1992), or failed retrieve and freely recall information about the feature when asked to provide a description (Lewicki et al., 1987; R. C. L. Lindsay et al., 1994; Meissner et al., 2007; Nelson, 1978). Indeed, at least some of our subjects who watched a distinctive culprit but failed to freely recall information about the feature behaved as though they recognised the feature during the lineup task; they made more identifications of the distinctive suspect in the unfair lineups than those subjects who had not had the opportunity to encode the feature (i.e., those subjects who had watched a non-distinctive culprit). Our data indicate that fair and unfair lineups result in equivalent discriminability when the culprit does not have the feature during the crime, but fair lineups significantly enhance people's ability to discriminate between innocent and guilty suspects more than unfair lineups when the culprit has the feature during the crime. Given that police officers can never be sure whether a witness who does not report a distinctive feature has no memory of that feature or has simply failed to report that feature, our results suggest that it is always sensible to construct fair lineups for suspects with distinctive features.

But how exactly should the police create fair lineups for distinctive suspects? Current police guidelines around the world provide conflicting advice about which lineup technique should be used when a suspect arrives for an identification parade with a distinctive feature that was not included in the witness's description (e.g., Police and Criminal Evidence Act 1984, Code D, 2011; Technical Working Group for Eyewitness Evidence, 2003). For the first time, we have shown that all three fair lineup techniques—replication, pixelation and block—are equally effective at enhancing subjects' ability to discriminate between innocent and guilty suspects, regardless of whether the culprit had the feature during the crime and the content of the witness's description. This has important implications for police practice,

because it suggests that the fair lineup techniques can be used interchangeably; the procedure could be selected due to specific requirements of a case, ease of application, or, in a time of austerity, even financial considerations.

Finally, our results also enhance our understanding of the conditions in which a culprit's change of appearance between the criminal event and the lineup task can affect identification performance. Perhaps somewhat surprisingly, we found that adding a distinctive feature or adding a block or pixelated area to conceal a distinctive feature between study (the crime) and test (the lineup) had no depreciable effect on identification accuracy or confidence judgements. Many face recognition and lineup studies have shown that changes to hairstyle, facial expression or pose can impair recognition accuracy (e.g., Ellis, 1975; Metzger, 2001; Patterson & Baddeley, 1977; Read, 1995; Shapiro & Penrod, 1986; Terry, 1994). Few eyewitness identification studies have, however, examined the effect of *adding* new features to lineup faces that were not present at encoding.[5] Nevertheless, our results are consistent with one eyewitness study in which an obvious stocking-mask disguise was added to all of the faces during the lineup task (Davies & Flin, 1984). Perhaps this suggests that witnesses are relatively immune to appearance changes, when the change involves the addition of a feature that was clearly not present during encoding. We also followed current police guidelines and informed our subjects that the lineup images may have been digitally altered in some way (A. Monaghan, National VIPER User Group, personal communication, August 15, 2016; C. Wilkinson, Northamptonshire police, personal communication, August 17, 2016). Given that the digital alterations were obvious, this instruction may have further prompted subjects to discount the feature or the pixelated or blocked out area and more carefully study the facial appearance of the lineup members (see also Porter, Moss, & Reisberg, 2014). It is important to note, however, that other studies have found that appearance-change instructions can harm identification accuracy and increase the number of false identifications without enhancing subjects' ability to

[5] In Experiment 3 of Read (1995), subjects attempted to identify a male whom they had encountered with no facial hair and no glasses from a lineup in which he (and the foils) had facial hair and glasses. Subjects also attempted to identify a female target whom they had encountered with glasses and her hair pulled back from a lineup in which she (and the foils) had no glasses and loose hair. Unfortunately, because subjects' identification responses were collapsed over the two targets, it is not possible to examine how the *addition* of features during the lineup test influenced identification accuracy.

correctly identify the culprit (Charman & Wells, 2007; Molinaro, Arndorfer, & Charman, 2013; Porter et al., 2014). Research is yet to examine the effect of instructions on eyewitness identification performance in lineups for distinctive suspects. This is certainly a necessary area for future research.

In sum, our findings are consistent with the diagnostic-feature-detection model suggesting that it remains a viable account of lineup decision-making. This is important because a well-tested and refined theory of lineup discriminability can ultimately guide the field to develop new procedures that further enhance eyewitness identification accuracy (Gronlund et al., 2015). Practically, our findings suggest that, when a suspect has a feature during the identification task that was not present during the crime, it does not seem to matter exactly *how* police officers prevent the distinctive suspect from standing out. Critically, though, it is possible for a witness to fail to recall a distinctive feature but then subsequently rely on it when presented with a lineup. Therefore, our study further echoes the necessity for fair lineup techniques for suspects with distinctive features, regardless of the content of the witness's description.

Chapter 5 :
Identification Performance on Replication Lineups

"…all the participants had different types of facial hair, some with mustaches, some with beards, goatees, etc. Nothing sets the picture of this defendant off by his facial hair."

People v. Adams (1982)

Overview

When constructing lineups for suspects with distinctive facial features there are three techniques—replication, pixelation and block—that police officers can use to create fair lineups by preventing the suspect from standing out. So far, our research has shown that, when the culprit has a distinctive feature during the crime, all three fair lineup techniques enhance people's ability to discriminate between innocent and guilty suspects compared to unfair (do-nothing) lineups in which the suspect is the only person with the distinctive feature. All three fair lineup techniques result in similar identification performance, and this is the case within different age groups (i.e., young, middle-aged and older adults) and regardless of the whether the culprit had the distinctive feature during the crime or the content of the witness's description of the culprit. Together, these findings support the diagnostic-feature-detection model (Wixted & Mickes, 2014) and suggest that there are three equally effective ways to foster accurate eyewitness identifications when creating fair lineups for distinctive suspects.

But might the way in which a lineup technique is applied alter how effective it is at fostering accurate eyewitness identifications? Block and pixelation techniques can be uniformly applied across the lineup members because the same black block is added to, or the same area is pixelated on, every face. There could, however, be more variability in how a suspect's feature is replicated across the lineup members. In this chapter, we examine how variation in the replicated feature across the foils affects eyewitness identification performance. In two experiments, we compared moderate-variation (Experiment 1, $N = 1{,}383$) and high-variation (Experiment 2, $N = 1{,}408$) lineups against low-variation and do-nothing lineups.

Introduction

If you had witnessed a crime committed by a man with a distinctive goatee, what would help to make your identification of the culprit more accurate: if you were presented with a police lineup in which the men all had the same facial hair, or if you were presented with a police lineup in which all the men had different types of facial hair? Approximately one third of all police suspects have a distinctive facial feature (e.g., a scar, tattoo, bruising; Flowe et al., 2010). In such cases, legal guidelines state that a suspect should not stand out in a lineup (Police and Criminal Evidence Act 1984, Code D, 2011; Technical Working Group for Eyewitness Evidence, 1999). Indeed, much empirical evidence has shown that leaving a suspect to stand out makes witnesses prone to identifying the suspect, regardless of whether that suspect is innocent or guilty (e.g., Doob & Kirshenbaum, 1973; Wells, Leippe, & Ostrom, 1979), and can harm witnesses' ability to tell the difference between innocent and guilty suspects (Chapters 2–4). To prevent suspects with distinctive features from standing out in lineups, police officers often digitally add the suspect's feature onto the other lineup members—the foils. Replicating a near-identical feature over the foils increases the number of guilty suspect (i.e., real culprit) identifications compared to simply removing the feature from the face of the suspect (Badham et al., 2013; Zarkadi et al., 2009) and enhances eyewitness identification accuracy more than leaving the distinctive suspect to stand out (Chapters 2–4). But how exactly should the police replicate a suspect's distinctive feature across the foils? In this chapter, we examine how variation in the replicated feature affects eyewitness identification performance and consider what this tells us, theoretically, about how eyewitnesses make identification decisions.

On the one hand, varying how the suspect's distinctive feature is replicated across the foils (*replication-with-variation*) might increase correct identifications of guilty suspects, without increasing incorrect identifications of innocent suspects. A distinctive facial feature is likely to be a salient cue that the witness will remember (Valentine, 1991; Winograd, 1981). When a feature is replicated with some variation, but remains within the constraints of the witness's description, the witness can rely on their memory of the distinctive feature to recognise the culprit (hereafter, the *replication-with-variation hypothesis*; Valentine, Hughes, & Munro, 2009). This *replication-with-variation hypothesis* accords with legal guidelines recommending

that a lineup should not consist of foils that are highly similar in appearance to the suspect—sometimes referred to as "clones" (Technical Working Group for Eyewitness Evidence, 2003). Indeed, we know that lineups containing foils that are very similar-looking to the suspect can impede correct identifications of guilty suspects (Fitzgerald, Oriet, & Price, 2015; Sauer, Brewer, & Weber, 2008; Wells et al., 1993). Therefore, replication-with-variation may enhance identification accuracy because it helps witnesses to identify the real culprit when he is in the lineup.

On the other hand, there is good reason to think that greater similarity across the lineup members should improve identification performance (hereafter, *replication-without-variation*). When witnesses compare similar faces in a lineup it provides them with useful information about which features should be used to make the identification decision (Wixted & Mickes, 2014). The diagnostic-feature-detection model proposes that there are some facial features that are non-diagnostic of guilt; these are features that are included in the witness's description, are shared by all lineup members, and, crucially, are shared by both innocent and guilty suspects alike. The model suggests that if witnesses rely on non-diagnostic features to make their identification decision, then their ability to tell the difference between innocent and guilty suspects will be impaired. A lineup composed of similar-looking individuals allows witnesses to immediately discount many features that are shared by all members and to focus on other features that are diagnostic of guilt (i.e., features that are not shared by both innocent and guilty suspects). Relying on diagnostic features, rather than non-diagnostic features, enhances witnesses' ability to discriminate between innocent and guilty suspects.

Indeed, research suggests that comparison of similar faces can enhance people's ability to discriminate between innocent and guilty suspects. For instance, presenting faces simultaneously benefits suspect discrimination accuracy more than presenting faces in isolation (Clark, 2012; Dobolyi & Dodson, 2013; Gronlund et al., 2012; Key et al., 2015; Mickes et al., 2012; Neuschatz et al., 2016; Wetmore et al., 2015). A growing body of evidence also suggests that greater similarity across lineup members can improve people's ability to discriminate between innocent and guilty suspects (e.g., Clark, 2012; Fitzgerald, Whiting, Therrien, & Price, 2014) and can facilitate correct identifications of guilty suspects (Gronlund et al., 2009). Even surrounding the suspect by the most similar-looking faces enhances discriminability

more than organising the same lineup members so that the suspect is surrounded by less similar-looking faces (Moreland, 2015).

Further evidence for the diagnostic-feature-detection account comes from research which has shown that replication-without-variation enhances subjects' ability to tell the difference between innocent and guilty suspects more than lineups in which the suspect (either innocent or guilty) was the only person with the distinctive feature—a "do-nothing" lineup (Chapters 2–4). Presumably, the replication lineups made it clear that the distinctive feature was shared by all members and so subjects discounted the feature when making their identification decision. Conversely, the do-nothing lineups did not make it clear that the distinctive feature should not be used when making the identification decision, and so subjects relied upon it, which impaired their ability to discriminate between innocent and guilty suspects because the feature was something that both the innocent and guilty suspect shared.

How then, according to the diagnostic-feature-detection account, might replication-with-variation influence people's ability to discriminate between innocent and guilty suspects? When there is little variation across the foils, then the feature should be discounted and ability to discriminate between innocent and guilty suspects should be enhanced compared to do-nothing lineups in which only the suspect has the distinctive feature. However, when there is greater variation across the foils, the distinctive feature is likely to be discounted to a lesser degree, and, as such, the expected improvement in discriminability would be less. In short, replication-without-variation should enhance performance more than replication-with-variation, and both should be better than doing nothing and leaving the distinctive suspect to stand out.

In sum, the replication-with-variation hypothesis suggests that greater variation in the feature across the foils will enhance people's ability to tell the difference between innocent and guilty suspects compared to when there is little variation in the feature across the foils. This is because greater variation in the feature will allow the witness to use their memory of the distinctive feature in the identification decision, which will increase the number of guilty suspect identifications, without increasing the number of innocent suspect identifications. However, to the extent that the distinctive feature is something that both the innocent and guilty suspect share, the

diagnostic-feature-detection model does not predict this pattern of results. The model predicts that little variation in the feature across the foils will enhance people's ability to tell the difference between innocent and guilty suspects compared to when there is greater variation in the feature across the foils. This is because little variation in the feature will lead to discounting of the (non-diagnostic) distinctive feature, thereby enhancing performance compared to doing nothing to prevent the suspect from standing out. Greater variation in the feature would lead to lesser discounting of that (non-diagnostic) feature, thereby enhancing ability to discriminate to a lesser degree.

To determine how variation in the feature across the foils affected witnesses' identification performance, we compared moderate-variation (Experiment 1) and high-variation (Experiment 2) lineups against low-variation and do-nothing lineups. Subjects watched a mock crime video, described the appearance of the culprit and then attempted to identify the culprit from a lineup. Subjects also reported their confidence in their identification decision.

Experiment 1

Method

Design

We used a 3 (lineup type: low-variation, moderate-variation, do-nothing) × 2 (target: present, absent) mixed design, with target manipulated within subjects. We recruited as many subjects as possible before the end of the winter term, aiming for at least 400 subjects with usable data in each of the three between-subject conditions.

Subjects

We recruited 1,443 subjects from social network sites who were entered into a prize draw for two £25 Amazon gift vouchers. All subjects completed the study online. We excluded 60 people (4%; between 16–26 subjects in each of the three between-subject conditions) who had experienced technical difficulties while watching the video (n = 11, < 1% in total), incorrectly answered an attention check question (n = 14, < 1% in total), or stated that they had seen one of the videos before or had completed the study more than once (n = 35, 2% in total). This resulted in a total sample size of 1,383: between 453 and 466 subjects in each of the three between-subject cells. Table 5.1 shows a demographic breakdown of the sample.

Table 5.1
Demographic Information for the Samples in Experiments 1 and 2

	Experiment 1	Experiment 2
Sex		
Male	258	227
Female	1,096	1,138
Other	3	0
Prefer not to say	26	43
Age (years)		
M	32.16	38.55
SD	14.03	13.61
Range	16–76	16–80
Prefer not to say	3	5
Race or ethnicity		
White or European	1,218	1,302
Latin or Hispanic	23	29
Black, African, or Caribbean	7	2
Asian	48	5
Mixed	28	12
Other	6	5
Prefer not to say	53	53

Materials

Videos

We used two 30 s mock crime videos from Chapters 2–4. The culprit in the *graffiti* video had a large bruise around his right eye and the culprit in the *mugging* video had a large tribal tattoo on his right cheek.

Lineups

We used 6-person simultaneous lineups that either contained the guilty suspect (i.e., the culprit) and five foils (a target-present lineup), or contained a designated innocent suspect and the same five foils (a target-absent lineup). In Chapter 2, we created a pool of 40 foils for each culprit. For each culprit in the current study, we randomly selected six foils from these pools, and then randomly selected one of these faces to serve as the designated innocent suspect.

For low-variation lineups, each foil (and the innocent suspect) had the culprit's distinctive feature added to them (see Figure 5.1b). Each feature was very similar in size, appearance and location to the culprit's distinctive feature, which reflects current police practice in several jurisdictions including England, Wales, New Zealand, Canada, and Germany. These foils were used in the replication lineups in the previous chapters. For moderate-variation lineups, each foil had a similar feature to the culprit added to them (Figure 5.1c). Each feature was different in size and appearance to the culprit's distinctive feature, but was in the same location. We used descriptions of the culprit's distinctive feature to create the distinctive features on the foils. After watching the mock crime video, subjects in Chapter 4 were given 2 min to describe the appearance of the culprit from memory. For each culprit, we randomly selected five descriptions that described the feature correctly with specific location and in detail (i.e., descriptions categorised as "code 5" in Chapter 4), and used one description for each foil face. For do-nothing lineups, the suspect's distinctive feature was visible and the foils had no distinctive features (Figure 5.1d).

These lineup construction strategies ensured that only the distinctive features on the foils varied across the low-variation, moderate-variation, and do-nothing conditions. Using the same foils in target-present and target-absent lineups ensured that the similarity between the witness's memory of the culprit and the foils was held constant in target-present and target-absent lineups. Using the same image of the culprit in target-present lineups ensured that the similarity between the culprit at the time of the crime and the culprit at the time of lineup was held constant across the three lineup conditions. Likewise, using the same innocent suspect in target-absent lineups ensured that the similarity between the culprit and the innocent suspect was held constant across the three lineup conditions. In an applied sense, this method reflects a worst-case scenario, because the innocent suspect had a very similar distinctive feature to the culprit. This method may overestimate the number of innocent suspect identifications compared to when the innocent suspect has a less similar feature to the culprit. Yet, overestimating the number of errors avoids the more serious issue of potentially underestimating the frequency of poor identification decisions in real investigations (e.g., Palmer et al., 2013).

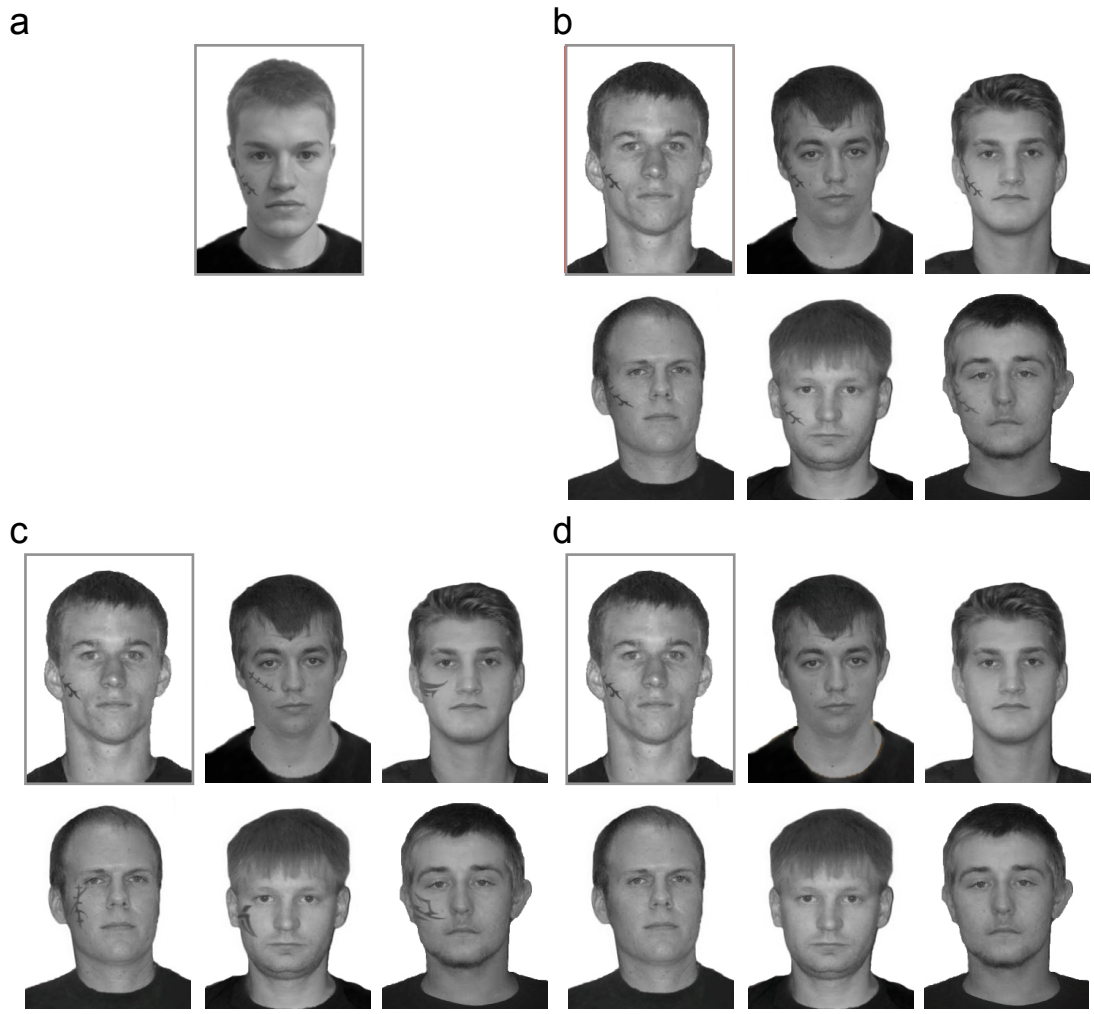

Figure 5.1. (a) A sample culprit, (b) a low-variation lineup, (c) a moderate-variation lineup, and (d) a do-nothing lineup in Experiment 1. Top left image in each lineup is the innocent suspect with a similar feature to the culprit.

Lineup fairness. We tested whether our lineup members were plausible alternatives to the culprit using images with no distinctive features. In Chapter 2, we formed a modal description of each culprit by asking 18 subjects to watch the mock crime video and then answer questions about the culprit's physical attributes (e.g., gender, eye colour, hair colour) from memory. We provided a group of mock witnesses with one modal description composed of eight descriptors, and then either a target-present or target-absent lineup. Forty different mock witnesses viewed each lineup (total $n = 160$) and were instructed to select the person in the lineup that best fit the description. Tredoux's E' uses the distribution of mock witness choices to determine how many plausible members are in the lineup (Tredoux, 1999). An effective size of 3 or higher suggests that a 6-person lineup is fair (Brigham, Ready, & Spier, 1990). For the *graffiti* scenario, Tredoux's E' was 4.37, 95% CI [3.53, 5.75]

for the target-present lineup and 3.92, 95% CI [3.12, 5.27] for the target-absent lineup. For the *mugging* scenario, Tredoux's *E′* was 4.17, 95% CI [3.47, 5.22] for the target-present lineup, and 3.79, 95% CI [3.22, 4.62] for the target-absent lineup. These results are similar to those found in previous archival (Valentine & Heaton, 1999) and laboratory (e.g., Horry et al., 2012) research and suggest that our lineup members fit the description of the culprit.

Manipulated images. To check that the distinctive features on our foils did not look doctored, we presented new subjects with two target-present lineups (n = 97). Subjects viewed one low-variation and one moderate-variation lineup, but did not see the same culprit twice. The position of the culprit in each lineup was random. We asked subjects to identify which photograph had *not* been digitally altered; they were no better than chance at this task in the low-variation graffiti and moderate-variation mugging lineups (ps > .22), and they were significantly worse than chance in the low-variation mugging and moderate-variation graffiti lineups (ps < .001). These findings suggest that our foil photos did not look digitally manipulated.

Similarity. To check that the distinctive features in our low-variation lineups were more similar to the suspect's feature than those in our moderate-variation lineups, we presented a new group of subjects (n = 26) with suspect-foil face pairs. For each pair, we asked subjects to rate how similar the distinctive features were on an 11-point Likert-type scale ranging from 0 (*not at all similar*) to 10 (*highly similar*). Subjects were instructed to pay attention to the distinctive feature (i.e., the black-eye or the tattoo). They were told: "We are interested in the similarity of a particular feature across the pairs of faces. We are not interested in the overall similarity of the two faces" and were encouraged to use the full range of the rating scale. Face-pairs were presented in two blocks, one for each mock crime scenario. The order of the blocks, and the order of the face-pairs within each block, was random. We used paired-samples t tests to assess the similarity ratings given to the same foil face in the low-variation and moderate-variation conditions. We also compared the average similarity rating given to the low-variation and moderate-variation foils (see Fitzgerald et al., 2015 for a similar approach). Table 5.2 shows that the distinctive features on the low-variation foils were rated as significantly more similar to the suspect's feature than those on the moderate-variation foils, in both target-present and target-absent conditions. The raters also rated the similarity

of the distinctive feature on the guilty and innocent suspect; the distinctive features on each guilty suspect-innocent suspect pair were highly similar (graffiti: $M = 9.27$, $SE = 0.15$; mugging: $M = 8.81$, $SE = 0.27$).

Table 5.2
Mean (SE) Ratings of Distinctive Feature Similarity in Experiment 1

		Variation condition			Cohen's *d* and 95% CI		
Suspect	Foil	Low	Moderate	*t*(25)	*d*	LL	UL
Graffiti							
Guilty	1	7.08 (0.31)	4.69 (0.42)	4.93**	1.26	0.84	1.86
	2	8.00 (0.31)	1.85 (0.38)	14.01**	3.47	2.39	4.53
	3	8.19 (0.23)	2.73 (0.42)	12.79**	3.18	2.17	4.18
	4	6.88 (0.37)	0.92 (0.24)	14.30**	3.72	2.57	4.85
	5	7.46 (0.25)	2.08 (0.37)	16.16**	3.36	2.34	4.37
	Average	7.28 (0.19)	2.45 (0.27)	20.10**	3.77	2.65	4.88
Innocent	1	6.65 (0.39)	4.73 (0.45)	3.70*	0.90	0.35	1.43
	2	7.77 (0.33)	1.85 (0.38)	12.77**	3.31	2.26	4.35
	3	7.92 (0.28)	2.69 (0.41)	12.35**	2.95	2.00	3.88
	4	7.08 (0.34)	1.31 (0.34)	15.99**	3.36	2.34	4.37
	5	7.00 (0.43)	2.62 (0.42)	8.28**	2.03	1.29	2.77
	Average	7.28 (0.23)	2.64 (0.30)	18.21**	3.40	2.39	4.40
Mugging							
Guilty	1	9.35 (0.91)	5.81 (0.41)	8.51**	2.16	1.38	2.93
	2	8.38 (0.29)	1.15 (0.30)	18.56**	4.76	3.35	6.15
	3	8.15 (0.35)	3.27 (0.46)	13.54**	2.33	1.60	3.04
	4	8.54 (0.27)	1.31 (0.32)	18.95**	4.79	3.37	6.19
	5	7.92 (0.36)	1.62 (0.360	16.45**	3.44	2.40	4.47
	Average	8.47 (0.23)	2.63 (0.28)	25.03**	4.49	3.20	5.77
Innocent	1	8.81 (0.28)	5.31 (0.46)	7.87**	1.80	1.12	2.46
	2	8.73 (0.28)	1.58 (0.36)	15.91**	4.38	3.05	5.69
	3	7.96 (0.34)	3.58 (0.36)	13.16**	2.43	1.66	3.18
	4	8.42 (0.35)	1.50 (0.36)	16.69**	3.84	2.68	4.98
	5	7.46 (0.32)	1.50 (0.33)	14.05**	3.58	2.46	4.67
	Average	8.28 (0.23)	2.69 (0.30)	20.57**	4.12	2.91	5.32

Note. Scale ranged from 0 (*not at all similar*) to 10 (*highly similar*). Cohen's *d* was estimated using the formula: $d = M_{diff} \div s_{av}$, where M_{diff} is the mean difference between the low-variation and moderate-variation foil, and s_{av} is the average of the standard deviations for the low-variation and moderate-variation foil (Cumming, 2012). CI = confidence interval; *LL* = lower limit; *UL* = upper limit.
*$p = .001$, **$p < .001$.

Procedure

Subjects were told that the study was about perception and memory and were randomly allocated into one of the three between-subject conditions (with the constraint that subject numbers were relatively equal in each condition). There were two main phases in the experiment. In the first phase, the heading “Video A” was present throughout. Subjects watched a mock crime video (either graffiti or mugging) and were told to watch the video carefully because they would be asked questions about it later. When the video had finished, subjects had the opportunity to report any technical difficulties they may have experienced while watching the video. Next, we gave them 2 min to type a description of the appearance of the male culprit in the video. We asked subjects to describe all of his different facial features and stated: “Unusual or distinctive features are particularly useful for the police. So please try and describe any unusual or distinctive features in as much detail as possible.” After this, we gave subjects a 4 min filler task in which they attempted spatial reasoning puzzles. Subjects then rated how confident they were that they would be able to recognise the culprit in “Video A” on an 11-point Likert-type scale (0% = *completely uncertain*, 100% = *completely certain*). Following this, subjects were told that they would be presented with a lineup which may or may not contain the culprit and they should click on the person that they believed was the culprit, or choose "Not Present" if they thought that the culprit was not in the lineup. The type of lineup viewed by the subject depended on the condition to which they had been randomly assigned (i.e., low-variation, moderate-variation, do-nothing) and was either target-present or target-absent. The lineup consisted of two rows of three photos presented simultaneously. After subjects had made their identification decision, they rated their confidence on an 11-point Likert-type scale (0% = *completely uncertain*, 100% = *completely certain*). To check that they had paid attention, we asked subjects what had happened in the video.

The second phase of the study then began and the heading “Video B” was present throughout. Subjects completed the same sequence of tasks as in phase one, but, this time, they viewed the alternative mock crime video (graffiti or mugging) and lineup format (target-present or target-absent). The order of the video and target conditions was counterbalanced. Finally, when phase two was complete, we asked

subjects a number of demographic questions and a question to ascertain if they had seen either of the videos before.

Results & Discussion

First, we checked the content of subjects' descriptions of the distinctive culprits. Then, to determine how variation in the replicated feature affected identification performance, we conducted ROC analysis, examined subjects' identification responses and fit a signal detection process model to our data (Wixted & Mickes, 2014). We gathered further information by examining subjects' ability to judge the accuracy of their suspect identification decisions.

Descriptions

Given that the replication-with-variation hypothesis suggests that the feature should be replicated within the constraints of the witness's description, we first examined the type and level of detail subjects freely recalled about the culprit's feature. We used the coding scheme from Chapter 4. Two coders, who were blind to purpose of the study, completed the coding independently. To assess interrater reliability, we randomly selected 5% of the descriptions to be coded by both coders and computed Siegel and Castellan's kappa: there was substantial agreement between the coders, $\kappa = 0.72$ (Landis & Koch, 1977). All coding discrepancies were resolved through discussion between the first author and the coders. Each coder then coded 50% of the remaining descriptions. Table 5.3 shows the frequencies of descriptions in each coding category; the majority described the feature correctly including specific location (code 3), but the remaining descriptions were more likely to contain fewer details (codes 1 and 2) than more details (codes 4 and 5). Interestingly, only a small proportion—around 8% of the descriptions provided—described the feature with the level of detail that we used to construct our moderate-variation lineups (code 5). This suggests that the majority of subjects did not describe the distinctive feature in great detail, despite having read instructions that encouraged them to provide as much information as possible.

Table 5.3
Percentages (and Frequencies) of Descriptions in Each Coding Category

Code	Experiment 1		Experiment 2	
0 = did not describe the feature	5.93	(164)	4.44	(125)
1 = described something to do with the feature	25.67	(710)	26.21	(738)
2 = described the feature correctly	16.74	(463)	15.09	(425)
3 = described the feature correctly *with specific location*	40.67	(1,125)	44.46	(1,252)
4 = described the feature correctly *in detail*	1.92	(53)	1.42	(40)
5 = described the feature correctly *with specific location and in detail*	7.95	(220)	7.28	(205)
6 = completed the task incorrectly	0.18	(5)	0.25	(7)
7 = did not write anything	0.94	(26)	0.85	(24)
Total	100.00	(2,766)	100.00	(2,816)

ROC analysis

Our ROC analysis measures subjects' ability to discriminate between guilty and innocent suspects, because both the replication-with-variation hypothesis (Valentine et al., 2009) and the diagnostic-feature-detection model (Wixted & Mickes, 2014) make predications about how variation in the feature across foils will influence identifications of suspects. We constructed our ROC curves and calculated our *p*AUC statistics in the same way as in Chapters 2–4. In each set of ROC analysis, we set the specificity (1 – FAR) using the FAR range covered by the least extensive curve.

Collapsed over the two mock crime scenarios. Figure 5.2 shows that, in the aggregate, the ROC curves for the low-variation and moderate-variation lineups lie on top of each other, while the do-nothing curve falls closer to the chance line. The ROC analysis showed that both low-variation and moderate-variation lineups were equally effective at enhancing subjects' ability to tell the difference between innocent and guilty suspects compared to when nothing was done to prevent the distinctive suspect from standing out. The *p*AUCs (specificity = .798) for both low-variation (*p*AUC = 0.055, 95% CI: 0.045, 0.066, $D = 3.75$, $p < .001$) and moderate-variation (*p*AUC = 0.057, 95% CI: 0.046, 0.068, $D = 3.75$, $p < .001$) lineups were significantly greater than the *p*AUC for do-nothing lineups (*p*AUC = 0.030, 95% CI: 0.023, 0.038). The *p*AUCs for low-variation and moderate-variation lineups were not significantly different, $D = 0.20$, $p = .84$. Figure 5.2 also shows that subjects were

more willing to identify the suspect in do-nothing lineups compared to low-variation and moderate-variation lineups, and more willing to identify the suspect in moderate-variation lineups compared to low-variation lineups. This is because the do-nothing curve extends further right than the low-variation and moderate-variation curves, and the moderate-variation curve extends slightly further right than the low-variation curve.

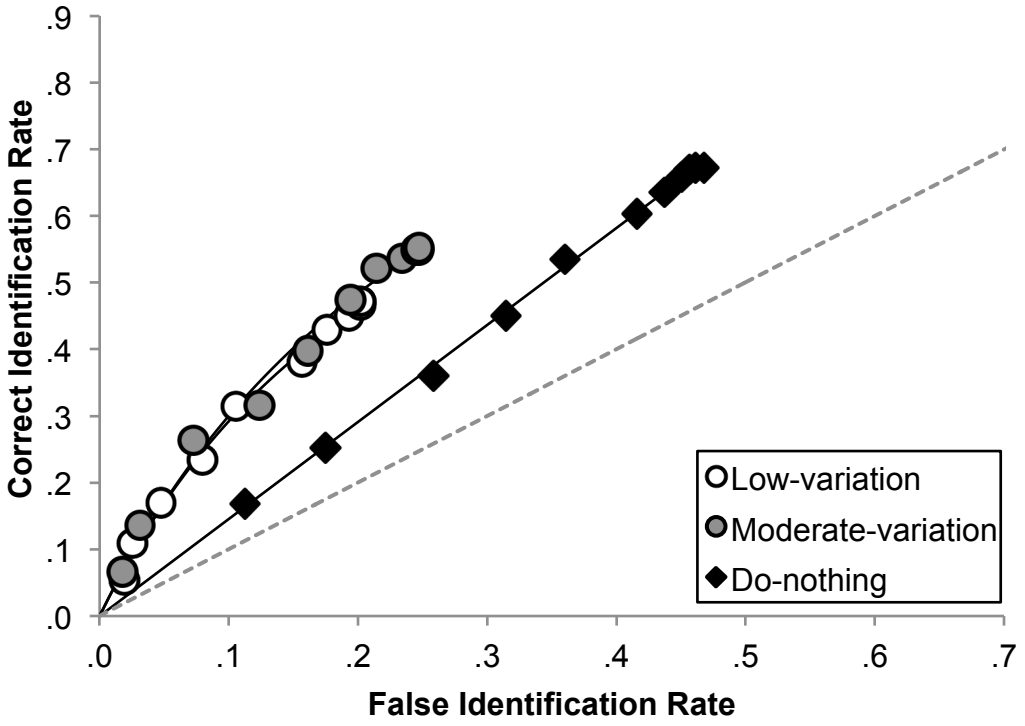

Figure 5.2. Receiver operating characteristic (ROC) curves for the low-variation, moderate-variation, and do-nothing lineups. Data are collapsed over the mugging and graffiti videos. The dashed line represents chance-level performance.

Separated by mock crime scenario. The lineups used in previous chapters were randomly generated for each subject using pools of faces, whereas, for each culprit in this study, we pre-designated one lineup member to be the innocent suspect and used the same set of five foil faces in target-present and target-absent lineups. While using set lineup members provides the greatest amount of experimental control, it restricts the variability in the test conditions (in our case, to only two guilty-innocent suspect pairs), which may limit the extent to which the findings can be generalized to other conditions (see, for instance, Brewer et al., 2010; D. S. Lindsay et al., 1998). One way to ensure that our findings are generalizable beyond a single guilty-innocent suspect pair is to check that the same pattern of results is

found in both stimulus sets. Therefore, we also analysed identification performance separately for each video.

Figure 5.3 shows the same pattern of results in both the (a) mugging and (b) graffiti videos: low-variation and moderate-variation lineups were equally effective at enhancing subjects' ability to tell the difference between innocent and guilty suspects compared to when nothing was done to prevent the distinctive suspect from standing out. In the mugging video, the *p*AUCs (specificity = .725) for both low-variation (*p*AUC = 0.078, 95% CI: 0.061, 0.098, $D = 2.36$, $p = .02$) and moderate-variation (*p*AUC = 0.083, 95% CI: 0.065, 0.101, $D = 6.69$, $p = .007$) lineups were significantly greater than the *p*AUC for do-nothing lineups (*p*AUC = 0.050, 95% CI: 0.037, 0.065). The *p*AUCs for low-variation and moderate-variation lineups were not significantly different, $D = 0.30$, $p = .77$. Similarly, in the graffiti video, the *p*AUCs (specificity = .871) for both low-variation (*p*AUC = 0.032, 95% CI: 0.021, 0.045, $D = 2.55$, $p = .01$) and moderate-variation (*p*AUC = 0.036, 95% CI: 0.025, 0.050, $D = 3.15$, $p = .002$) lineups were significantly greater than the *p*AUC for do-nothing lineups (*p*AUC = 0.013, 95% CI: 0.008, 0.022). The *p*AUCs for low-variation and moderate-variation lineups were not significantly different, $D = 0.53$, $p = .60$. Furthermore, in both videos, subjects were more willing to identify the suspect in do-nothing lineups compared to low-variation and moderate-variation lineups, and more willing to identify the suspect in moderate-variation lineups compared to low-variation lineups, though the difference between the low- and moderate-variation lineups was only very small in the graffiti video.

In sum, the ROC analysis illustrates that ability to discriminate between innocent and guilty suspects was similar in both low-variation and moderate-variation lineups, and, as such, does not provide conclusive support for either the replication-with-variation or the diagnostic-feature-detection accounts. However, both replication (low-variation and moderate-variation) lineups enhanced subjects' ability to discriminate between innocent and guilty suspects more than doing nothing—an effect that is predicted by the diagnostic-feature-detection model. The ROC analysis also indicated that greater variation in the feature across the foils (from low-variation to do-nothing) increased subjects' willingness to identify the guilty or innocent suspect. These findings were the same in both mock crime scenarios, suggesting that the results generalize beyond one guilty-innocent suspect

pair. We were, however, primarily interested in examining general effects, rather than comparing the idiosyncratic differences in identification performance across different distinctive features or criminal events. Therefore, for the subsequent analyses in Experiment 1, we collapsed our data over both videos.

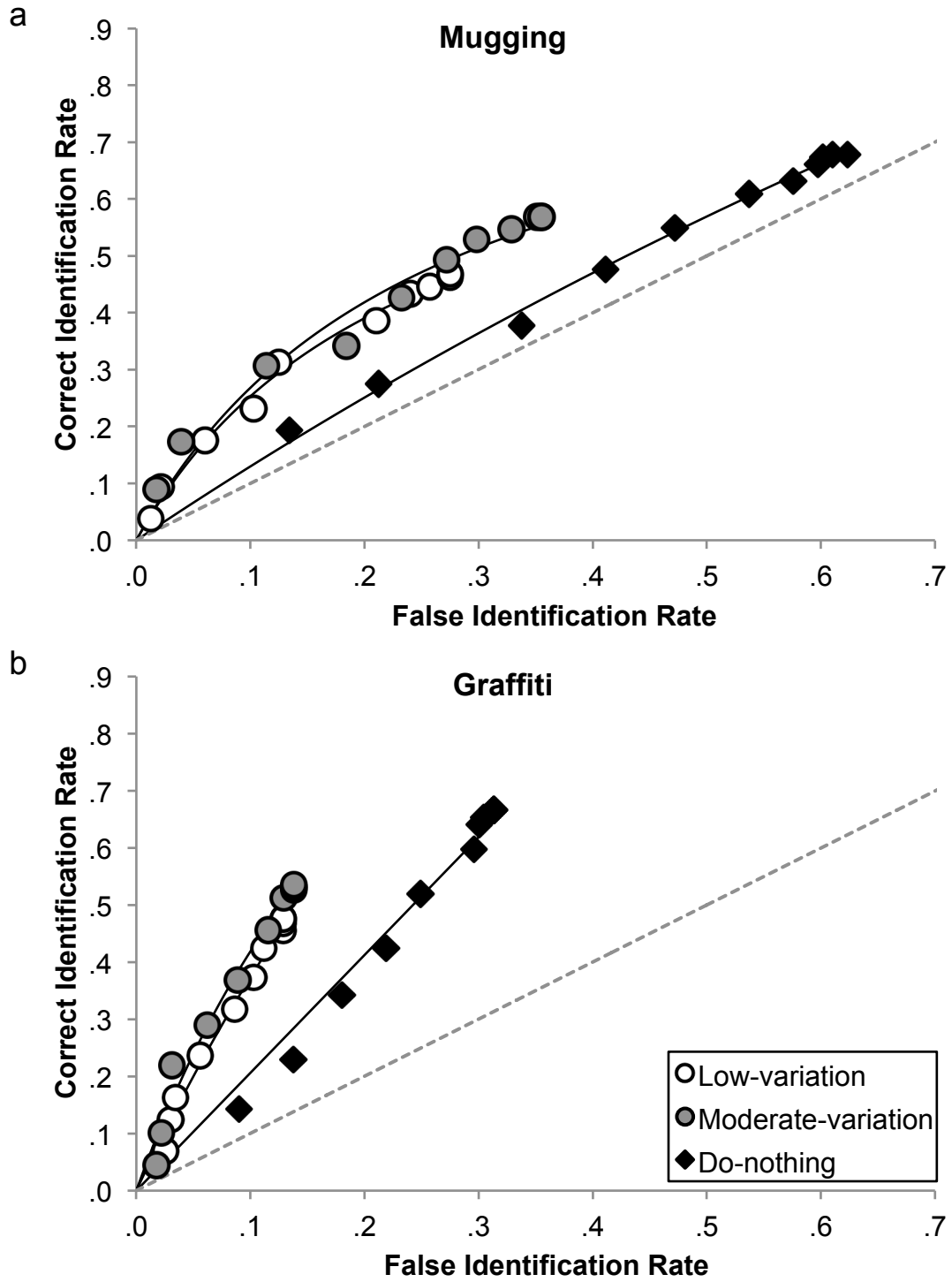

Figure 5.3. Receiver operating characteristic (ROC) curves for the low-variation, moderate-variation, and do-nothing lineups in the (a) mugging and (b) graffiti videos. The dashed lines represent chance-level performance.

Identification responses

Increasing the variation in the feature from low to moderate had no discernable impact on subjects' ability to tell the difference between innocent and guilty suspects. To investigate whether low-variation and moderate-variation lineups result in a similar pattern of identification responses, we calculated the proportion of suspect identifications, foil identifications and lineup rejections (i.e., "Not Present" responses) in each lineup type. Figure 5.4 shows the identification responses made in low-variation, moderate-variation, and do-nothing (a) target-present and (b) target-absent lineups.

Target-present lineups. Figure 5.4 a shows that greater variation in the distinctive feature across the lineup members resulted in more guilty suspect identifications. A 3 (lineup type: low-variation, moderate-variation, do-nothing) × 3 (identification response: guilty suspect, foil, incorrect rejection) two-way chi-square analysis indicated that lineup technique influenced identification responses, χ^2 (4, N = 1,383) = 41.49, $p < .001$, Contingency Coefficient $C = .17$. Low-variation lineups resulted in fewer guilty suspect IDs ($z = –2.68$, $p < .01$) and more foil IDs ($z = 2.83$, $p < .01$), while do-nothing lineups resulted in more guilty suspect IDs ($z = 3.06$, $p < .01$) and fewer foil IDs ($z = –2.13$, $p < .05$) and fewer lineup rejections ($z = –2.78$, $p < .01$) than expected. Specifically, three 2 (lineup type) × 2 (identification response: guilty suspect, foil) two-way chi-square analyses indicated that when subjects made a selection, subjects who viewed the do-nothing lineup were 1.46 times more likely to identify the guilty suspect than subjects who viewed the moderate-variation lineup, χ^2 (1, N = 703) = 4.02, $p = .04$, OR = 1.46, 95% CI [0.99, 2.15], and 2.42 times more likely to identify the guilty suspect than subjects who viewed the low-variation lineup, χ^2 (1, N = 708) = 25.05, $p < .001$, OR = 2.42, 95% CI [1.68, 3.50]. Subjects who viewed the moderate-variation lineup were 1.66 times more likely to identify the guilty suspect than subjects who viewed the low-variation lineup, χ^2 (1, N = 657) = 8.43, $p = .004$, OR = 1.66, 95% CI [1.16, 2.38].

Target-absent lineups. Figure 5.4b shows that greater variation in the distinctive feature across the lineup members resulted in more innocent suspect identifications. A 3 (lineup type: low-variation, moderate-variation, do-nothing) × 3 (identification response: innocent suspect, foil, correct rejection) two-way chi-square analysis indicated that lineup technique influenced identification responses, χ^2 (4, N

= 1,383) = 105.59, $p < .001$, Contingency Coefficient $C = .27$. Low-variation lineups resulted in fewer innocent suspect IDs ($z = –4.07$, $p < .001$) and more foil IDs ($z = 4.10$, $p < .001$), while moderate-variation lineups resulted in fewer innocent suspect IDs ($z = –2.26$, $p < .05$) and more lineup rejections ($z = 2.91$, $p < .01$) than expected. Do-nothing lineups resulted in more innocent suspect IDs ($z = 6.30$, $p < .001$) and fewer foil IDs ($z = –2.97$, $p < .01$) and fewer lineup rejections ($z = –2.97$, $p < .01$) than expected. Specifically, three 2 (lineup type) × 2 (identification response: innocent suspect, foil) two-way chi-square analyses indicated that when subjects made a selection, subjects who viewed the do-nothing lineup were 2.29 times more likely to identify the innocent suspect than subjects who viewed the moderate-variation lineup, χ^2 (1, $N = 543$) = 21.64, $p < .001$, OR = 2.29, 95% CI [1.59, 3.31], and 4.23 times more likely to identify the innocent suspect than subjects who viewed the low-variation lineup, χ^2 (1, $N = 589$) = 70.32, $p < .001$, OR = 4.23, 95% CI [2.96, 6.08]. Subjects who viewed the moderate-variation lineup were 1.85 times more likely to identify the innocent suspect than subjects who viewed the low-variation lineup, χ^2 (1, $N = 502$) = 11.29, $p < .001$, OR = 1.85, 95% CI [1.27, 2.69].

Although moderate-variation lineups resulted in more innocent suspect identifications than low-variation lineups, Figure 5.4b also shows that moderate-variation lineups resulted in the greatest number of correct decisions—lineup rejections. Three 2 (lineup type) × 2 (identification response: correct, incorrect) two-way chi-square analyses indicated that subjects who viewed the moderate-variation lineup were 1.41 times more likely to correctly reject the lineup than subjects who viewed the low-variation lineup, χ^2 (1, $N = 919$) = 6.64, $p = .01$, OR = 1.41, 95% CI [1.08, 1.84], and 2.08 times more likely to correctly reject the lineup than subjects who viewed the do-nothing lineup, χ^2 (1, $N = 917$) = 29.26, $p < .001$, OR = 2.08, 95% CI [1.58, 2.75]. Subjects who viewed the low-variation lineup were 1.48 times more likely to correctly reject the lineup than subjects who viewed the do-nothing lineup, χ^2 (1, $N = 930$) = 8.27, $p = .004$, OR = 1.48, 95% CI [1.12, 1.96].

In sum, our analyses of the identification responses align with the findings of the ROC analysis and suggest that greater variation in the feature across the foils (from low-variation to do-nothing) increased subjects' willingness to identify the guilty or innocent suspect. However, we also found that low-variation led to more foil identifications in both target-present and target-absent lineups than moderate-

variation and do-nothing lineups, while moderate-variation led to the greatest number of correct rejections in target-absent lineups.

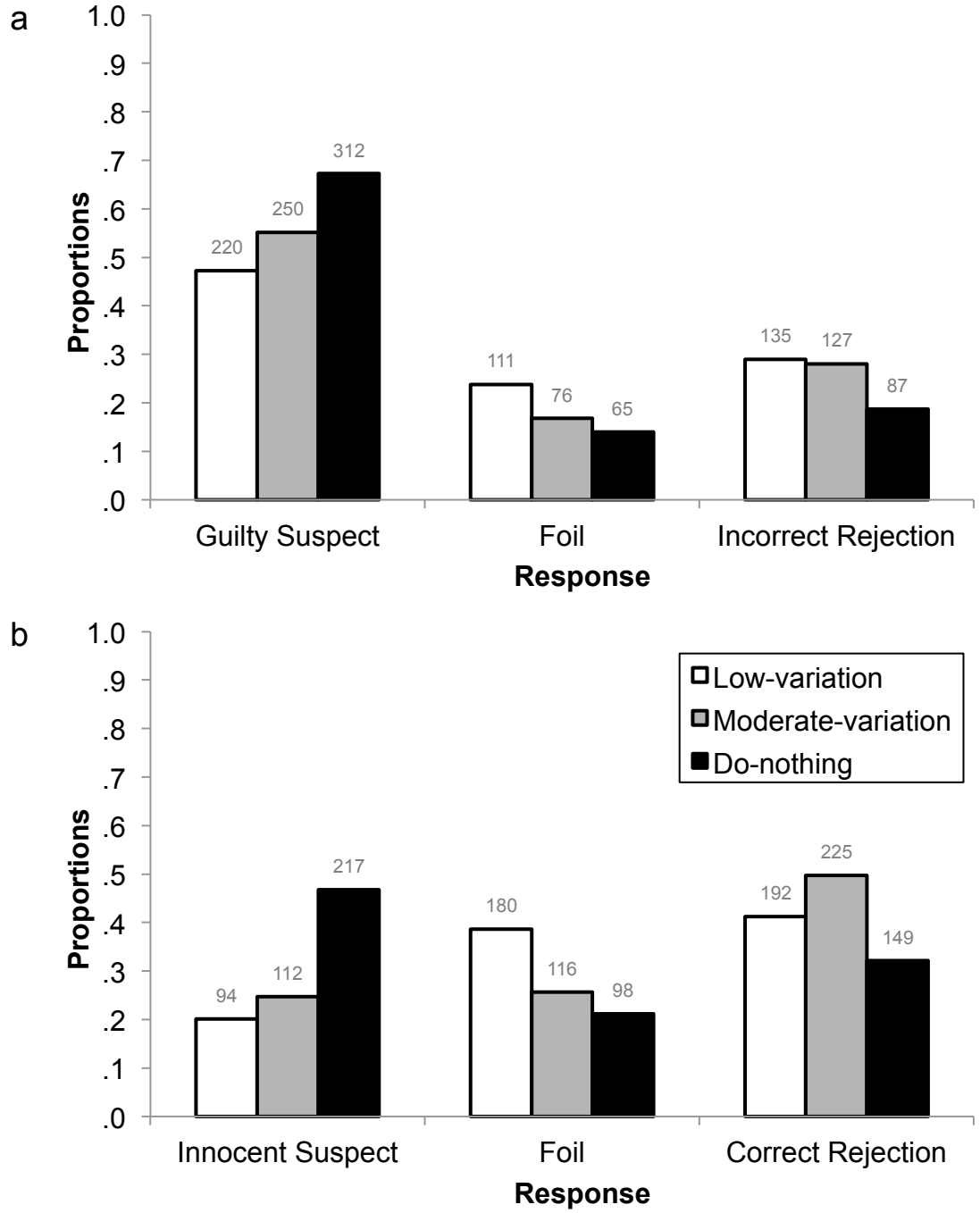

Figure 5.4. Identification responses made in low-variation, moderate-variation, and do-nothing (a) target-present and (b) target-absent lineups. Data labels are absolute frequencies.

Modelling

Low and moderate-variation lineups were equally effective at enhancing subjects' ability to discriminate between innocent and guilty suspects in the ROC analysis, but led to different patterns of foil identifications and lineup rejections in the identification response analysis. To investigate this further, we fit a signal detection model to our data (Wixted & Mickes, 2014; see Chapter 1 for a description of the model). Recall that for lineups in which the innocent suspect is more similar to the culprit than the other foils, the model consists of three memory strength distributions (μ_{guilty}, $\mu_{innocent}$, and μ_{foil}). The distance between the μ_{guilty} and $\mu_{innocent}$ distributions (d') measures subjects' ability to discriminate guilty suspects from innocent suspects and corresponds to the discriminability measures in our ROC analysis. Similarly, the distance between the μ_{guilty} and μ_{foil} distributions measures subjects' ability to discriminate guilty suspects from foils, and the distance between the $\mu_{innocent}$ and μ_{foil} distributions measures subjects' ability to discriminate innocent suspects from foils.

The model also assumes that there is a set of response criteria that reflect different levels of confidence. To limit the number of parameters, we collapsed our data down to a 5-point confidence scale: 0–20 (c_1), 30–40 (c_2), 50–60 (c_3), 70–80 (c_4), and 90–100 (c_5). These intervals resulted in a relatively similar number of identification decisions at each confidence level in each lineup condition. To reiterate, the model assumes that when a face is familiar enough to exceed the lowest criterion (c_1) then an identification is made; if more than one face exceeds c_1, then the face with the highest familiarity value is identified. The confidence in the identification is determined by the highest criterion that is exceeded.

The data contained 20 degrees of freedom, corresponding to the 5 levels of confidence for guilty suspect identifications and foil identifications in target-present lineups, and the 5 levels of confidence for innocent suspect identifications and foil identifications in target-absent lineups. Once these frequencies were known, the number of rejections made in target-present and target-absent lineups was fixed. We fixed $\mu_{innocent}$ to 0 and set the standard deviations for each distribution to 1 (i.e., we used an equal-variance model), for simplicity. Thus, the model for each lineup had 7 free parameters (μ_{guilty}, μ_{foil}, c_1, c_2, c_3, c_4, c_5) and the fit had 20 – 7 = 13 degrees of freedom.

We fit the model to our low-variation, moderate-variation and do-nothing lineup data by minimising the chi-square goodness-of-fit statistic. Our observed data and the values predicted by the best-fitting equal-variance model are shown in Table 5.4, and the best-fitting parameters and the chi-square goodness-of-fit statistics are shown in Table 5.5. While the simple equal-variance model captured the trends in our data (see Table 5.4), the significant chi-square goodness-of-fit statistics in the left-hand column (full model) of Table 5.5 indicate that our data deviated from the predictions of this simple model, suggesting that a more complex model might fit the data better.[6] Figure 5.5 displays the parameters estimated by the best-fitting equal-variance model for the three lineup types. Before we turn to our main measure of interest—*d′*—it is immediately obvious that the distance between the innocent suspect distribution and the guilty suspect distribution is smaller than the distance between the foil distribution and the guilty suspect distribution, even in the low-variation lineups where all of the lineup members had a very similar feature. This indicates that one (or perhaps both) of our designated innocent suspects was more similar to the guilty suspect than were the other foils. It was not our intention to select innocent suspects who were relatively more similar to the guilty suspects. Nevertheless, because we used the same designated innocent suspects in all three lineups conditions, the similarity between the innocent and guilty suspects was held constant across our lineup manipulation. As such, our study still provides a valid test of the influence of the variation in the feature across the foils.

Now, returning to our main measure of interest—*d′*. Figure 5.5 shows that *d′* is considerably larger in both low-variation and moderate-variation lineups than do-nothing lineups. Interestingly, the model also estimates that *d′* is larger in moderate-variation lineups than low-variation lineups. To test whether the observed differences in *d′* were statistically significant, we performed three pairwise comparisons: low-variation versus moderate-variation, low-variation versus do-nothing, and moderate-variation versus do-nothing. We fit the same model, allowing the confidence criteria to differ, but constraining *d′* to be equal in the two lineups being compared. The overall χ^2, df and *p* rows in Table 5.5 show the full

[6] When we fit the data to the low-variation, moderate-variation and do-nothing lineup data separately, the model fit the low-variation data well ($p = .15$), but significantly deviated from the observed data in the moderate-variation ($p = .04$) and do-nothing ($p = .03$) conditions.

(unconstrained) and constrained model fit statistics. In comparison to the full model, the constrained model provided a significantly worse fit of the data for the low-variation and moderate-variation, χ^2 (2) = 22.25, $p < .001$, low-variation and do-nothing, χ^2 (2) = 125.84, $p < .001$, and moderate-variation and do-nothing, χ^2 (2) = 51.31, $p < .001$, lineup comparisons. These results indicate that the differences in d' across the low-variation, moderate-variation and do-nothing lineups were statistically significant.

It should be noted that our moderate-variation and do-nothing data significantly deviated from the predictions of the equal-variance model. The relatively poor fit may simply reflect high power to detect even slight deviations from the simple model as a result of our large *N* (Colloff, Wade, & Strange, 2016, supplemental materials). The model fit in the do-nothing condition—but not the low-variation or moderate-variation conditions—was significantly improved by allowing for unequal variance. When we fit an unequal-variance model to our data, we found the same results. In sum, the results of the model fitting are consistent with our findings of the ROC analysis and suggest that both low-variation and moderate-variation lineups enhance discriminability more than doing nothing to prevent the distinctive suspect from standing out. The model-fitting exercise, however, suggested that moderate-variation lineups enhanced people's ability to discriminate between innocent and guilty suspects more than low-variation lineups, whereas the ROC analysis indicated that both were equally effective.

What could explain the different results found in the modelling and the ROC analysis? If we look back at the best-fitting parameters in Figure 5.5, it is clear that the placement of the foil distribution changes across the different lineup types. The model finds the best-fitting parameters to accommodate all three distributions (i.e., identifications of foils, innocent suspects and guilty suspects). Because low-variation lineups resulted in a greater proportion of foil identifications, the model required a greater overlap of the foil and both (guilty and innocent) suspect distributions in the low-variation than the moderate-variation lineups. It is likely that this can account for the smaller distance between the innocent and guilty distributions (i.e., the smaller d') in the low-variation lineups, compared to the moderate-variation lineups. Indeed, when we fit the same model but discounted foil identifications, as expected,

there was no statistically significant difference in d' across the low-variation and moderate-variation lineup conditions.

The finding that there was greatest overlap of the foil and suspect distributions in the low-variation lineup illustrates that subjects more easily confused foils with suspects in low-variation lineups than in moderate-variation and do-nothing lineups. This finding makes good sense. In low-variation lineups it is likely that all of the foils seem familiar (i.e., evoke some memory signal) because they all have a feature that is very similar to the culprit's. In moderate-variation and do-nothing lineups, however, the foils seem less familiar (i.e., evoke a less strong memory signal) because they each have a feature that is only moderately similar to the culprit's, or they have no feature at all. Thus, the difference between the memory signals evoked by the foils and the suspects is smaller in the low-variation lineups than in the moderate-variation lineups, and markedly smaller in the low-variation lineups than in the do-nothing lineups.

Table 5.4

Observed and Predicted Identification Responses in Each Confidence Bin in the Low-variation, Moderate-variation, and Do-nothing Lineups in Experiment 1

	Low-variation						Moderate-variation						Do-nothing					
	Target present			Target absent			Target present			Target absent			Target present			Target absent		
Confidence	Guilty suspect	Foil	Inc. reject	Innocent suspect	Foil	Correct reject	Guilty suspect	Foil	Inc. reject	Innocent suspect	Foil	Correct reject	Guilty suspect	Foil	Inc. reject	Innocent suspect	Foil	Correct reject
0–20																		
Observed	10.00	8.00	-	4.00	12.00	-	7.00	3.00	-	6.00	14.00	-	7.00	5.00	-	8.00	13.00	-
Predicted	7.26	8.62	-	5.76	13.14	-	8.09	6.16	-	6.97	10.37	-	9.31	5.73	-	10.56	8.66	-
30–40																		
Observed	33.00	28.00	-	17.00	35.00	-	28.00	26.00	-	18.00	25.00	-	25.00	20.00	-	16.00	25.00	-
Predicted	26.01	27.92	-	18.72	40.45	-	28.25	18.97	-	21.95	30.12	-	26.93	14.30	-	28.80	20.75	-
50–60																		
Observed	68.00	39.00	-	36.00	63.00	-	72.00	24.00	-	32.00	39.00	-	71.00	23.00	-	47.00	30.00	-
Predicted	57.66	48.18	-	33.38	63.14	-	59.07	29.34	-	36.71	41.69	-	61.34	23.33	-	58.15	31.45	-
70–80																		
Observed	58.00	24.00	-	5.00	48.00	-	81.00	17.00	-	42.00	24.00	-	92.00	13.00	-	65.00	20.00	-
Predicted	58.47	32.10	-	24.71	37.50	-	76.51	21.93	-	33.41	27.08	-	83.21	17.86	-	66.00	21.88	-
90–100																		
Observed	51.00	12.00	-	12.00	22.00	-	62.00	6.00	-	14.00	14.00	-	117.00	4.00	-	81.00	10.00	-
Predicted	56.25	14.24	-	14.37	15.14	-	68.52	6.89	-	16.56	7.49	-	122.55	8.11	-	69.20	8.97	-
Total																		
Observed	-	-	135.00	-	-	192.00	-	-	127.00	-	-	225.00	-	-	87.00	-	-	149.00
Predicted	-	-	129.30	-	-	199.70	-	-	129.26	-	-	220.66	-	-	91.32	-	-	139.57

Note. The total row displays all reject identification decisions because the model does not account for the confidence level with which lineup rejections are made. Inc. reject = incorrect rejection; Correct reject = correct rejection.

Table 5.5

Full and Constrained (d') Model Fits for the Low-variation vs. Moderate-variation, Low-variation vs. Do-nothing, and Moderate-variation vs. Do-nothing Comparisons in Experiment 1

	Full model		Constrained model	
Estimate	Low-variation	Moderate-variation	Low-variation	Moderate-variation
μ_{guilty} (d')	0.70	0.76	0.73	0.73
μ_{foil}	–0.61	–0.92	–0.76	–0.76
c_1	0.65	0.54	0.55	0.63
c_2	0.71	0.61	0.61	0.69
c_3	0.92	0.81	0.83	0.89
c_4	1.35	1.20	1.25	1.28
c_5	1.86	1.79	1.76	1.86
Overall χ^2	40.97		63.22	
Overall df	26		28	
Overall p	.03		< .001	
	Low-variation	Do-nothing	Low-variation	Do-nothing
μ_{guilty} (d')	0.70	0.41	0.47	0.47
μ_{foil}	–0.61	–1.59	–1.13	–1.13
c_1	0.65	–0.15	0.24	0.11
c_2	0.71	–0.06	0.31	0.19
c_3	0.92	0.14	0.52	0.37
c_4	1.35	0.53	0.94	0.73
c_5	1.86	1.04	1.45	1.21
Overall χ^2	42.08		63.22	
Overall df	26		28	
Overall p	.02		<.001	
	Moderate-variation	Do-nothing	Moderate-variation	Do-nothing
μ_{guilty} (d')	0.76	0.41	0.53	0.53
μ_{foil}	–0.92	–1.59	–1.28	–1.28
c_1	0.54	-0.15	0.27	0.05
c_2	0.61	–0.06	0.34	0.13
c_3	0.81	0.14	0.54	0.32
c_4	1.20	0.53	0.94	0.69
c_5	1.79	1.04	1.50	1.18
Overall χ^2	46.71		98.02	
Overall df	26		28	
Overall p	.008		<.001	

Note. The full model allows d' to differ between the two lineups being compared. The constrained model holds d' constant across the two lineups being compared. Overall χ^2, df and p rows represent goodness-of-fit statistics when the model was fit to the two lineups together.

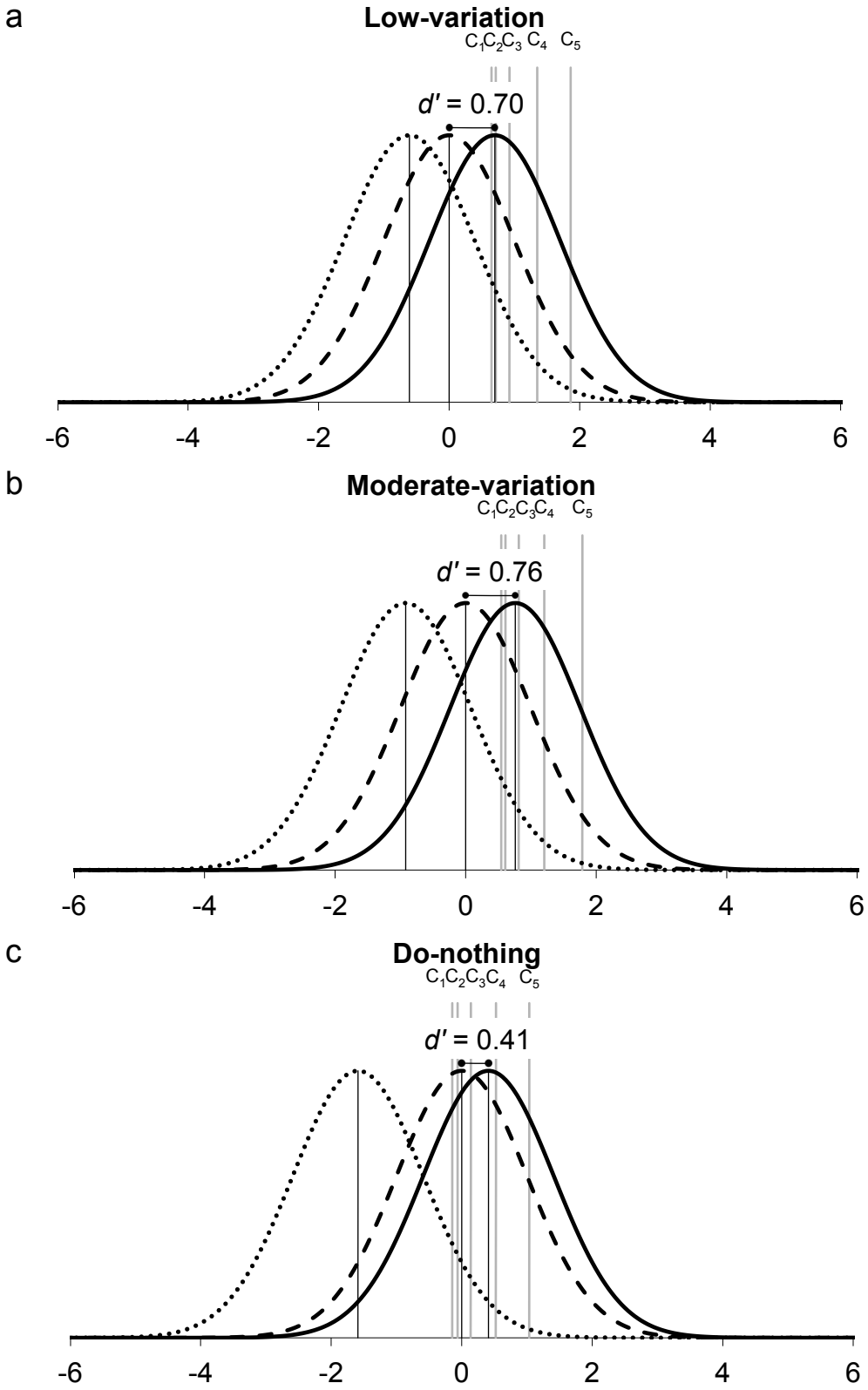

Figure 5.5. Foil, innocent suspect, and guilty suspect distributions for (a) low-variation, (b) moderate-variation, and (c) do-nothing lineups using the best-fitting equal-variance signal detection model parameters. *d′* measures subjects' ability to discriminate between innocent and guilty suspects. c_1, c_2, c_3, c_4 and c_5 are a set of response criteria that reflect different levels of confidence.

Confidence and accuracy

Previous research shows subjects who are presented with an unfair do-nothing lineup make less accurate suspect identifications at every level of confidence than subjects who are presented with a fair replication-without-variation lineup (Chapters 2–4). According to the diagnostic-feature-detection account, this occurs because it is not clear to the witness that the suspect's distinctive feature is unhelpful in do-nothing lineups and, as such, subjects fail to lower their confidence judgement despite using the (non-diagnostic) feature to make their identification decision. In the current study, moderate-variation lineups resulted in more suspect identifications than low-variation lineups, perhaps suggesting that the distinctive feature is relied upon to a greater degree in moderate-variation lineups. If subjects viewing moderate-variation lineups fail to lower their confidence judgement, despite relying on the (non-diagnostic) feature to a greater degree, then they will be less accurate at a particular level of confidence than subjects who have viewed a low-variation lineup.

To test this, we constructed confidence-accuracy curves in the same way as in Chapters 2–4, but, this time, we collapsed the confidence ratings to three categories to provide more stable estimates (0–60%, 70–80%, 90–100%, see Mickes, 2015). These categories ensured a relatively similar number of identification decisions at each confidence level in each lineup type. The frequencies of identification responses in each confidence bin are presented in Appendix I. Figure 5.6 shows the confidence-accuracy curves for each lineup technique. Nonoverlapping error bars signify reliable differences in the proportion of correct suspect identifications (e.g. Sauer et al., 2010). Descriptively speaking, both low-variation and moderate-variation lineups resulted in more accurate suspect identifications than do-nothing lineups, but only low-variation lineups resulted in significantly more accurate suspect identifications than do-nothing lineups at every confidence level. Moderate-variation lineups resulted in significantly more accurate suspect identifications than do-nothing lineups at both the lower (i.e., 0–60% certain) and higher (i.e., 90–100% certain) end of the confidence scale. These findings fit with the diagnostic-feature-detection account: It was not clear that the feature was unhelpful in the do-nothing lineups, and therefore subjects failed to make more conservative confidence judgements, despite relying on the feature to make their identification decision.

Low-variation lineups resulted in more accurate suspect identifications than moderate-variation lineups, but only significantly so at lower levels of confidence (i.e., 0–60% certain). Figure 5.6 shows that both low- and moderate-variation lineups resulted in a similar proportion of correct suspect identification at the highest level of confidence (i.e., 90–100% certain). Theoretically, this might suggest that subjects who made low-confidence suspect identifications from a moderate-variation lineup relied on the feature to a greater extent than those who made low-confidence identifications from a low-variation lineup and those who viewed the moderate-variation lineup failed decrease their confidence judgements to the degree required to account for their reduced accuracy. It might also suggest that subjects who made highly confident suspect identifications discounted the feature to the same extent in both low- and moderate-variation lineups and, as such, identifications were equally likely to be accurate in both lineup types. In sum, high confidence identifications were equally likely to be accurate on low-variation and moderate-variation lineups, but, generally speaking, low-variation lineups seem to be the most effective way of enhancing the accuracy of suspect identifications across the range of confidence ratings.

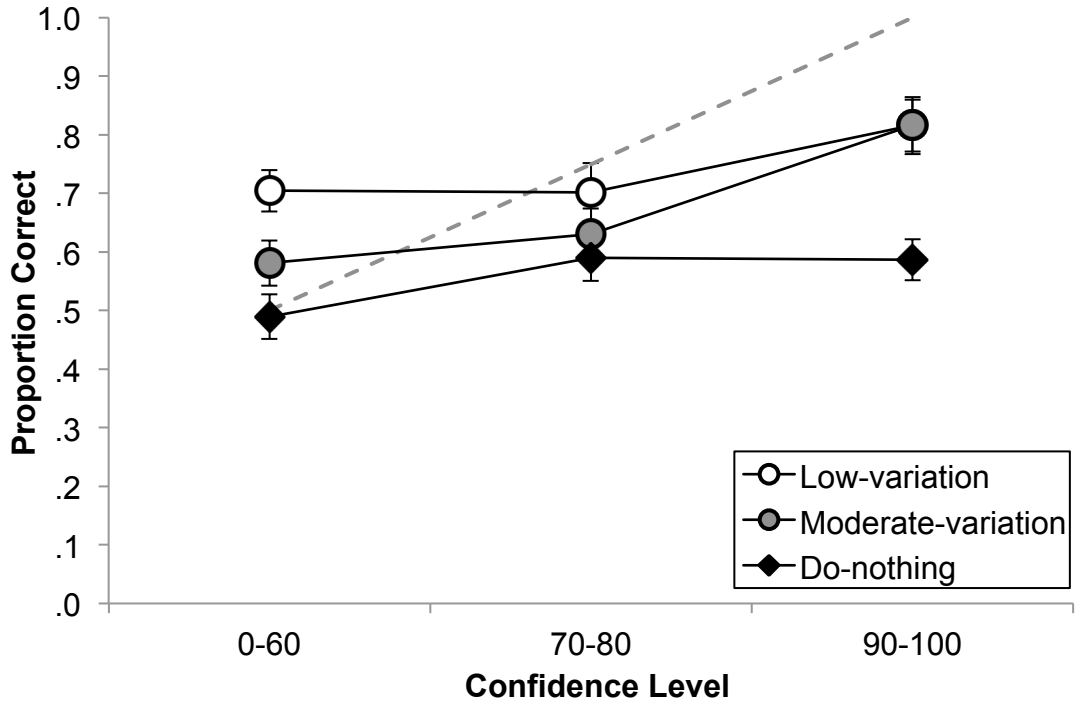

Figure 5.6. Confidence-accuracy curves for suspect identifications in the low-variation, moderate-variation, and do-nothing lineups. Error bars indicate ±1 SE. The dashed line represents chance accuracy at the lowest confidence bin (i.e., 0–60) and perfect accuracy at the highest confidence bin (i.e., 90–100).

Experiment 2

In Experiment 1, we examined how variation in the replicated feature affects eyewitness identifications by comparing performance in low-variation, moderate-variation and do-nothing lineups. The replication-with-variation hypothesis suggests that greater variation in the feature across the foils will enhance people's ability to tell the difference between innocent and guilty suspects compared to when there is little variation in the feature across the foils. This is because greater variation in the feature will allow the witness to use their memory of the distinctive feature in the identification decision, which will increase the number of guilty suspect identifications, without increasing the number of innocent suspect identifications (Valentine et al., 2009). Conversely, the diagnostic-feature-detection model predicts that little variation in the feature across the foils will enhance people's ability to tell the difference between innocent and guilty suspects compared to when there is greater variation in the feature across the foils. This is because little variation in the feature will lead to discounting of the (non-diagnostic) distinctive feature, thereby enhancing performance compared to doing nothing to prevent the suspect from standing out. Greater variation in the feature would lead to lesser discounting of that (non-diagnostic) feature, thereby enhancing ability to discriminate to a lesser degree.

In Experiment 1, our ROC analysis showed that ability to discriminate between innocent and guilty suspects was similar in both the low- and moderate-variation lineups, which is not predicted by either the replication-with-variation hypothesis or the diagnostic-feature-detection model. There were, however, some benefits of moderate-variation lineups. Subjects who viewed low-variation lineups were more likely to shift their identification from the (guilty or innocent) suspect to another foil, whereas subjects who viewed moderate-variation lineups either tended to identify the (guilty or innocent) suspect or correctly reject the lineup when the real culprit was not present. It is important to note, however, that these are not the benefits predicted by the replication-with-variation hypothesis. Our ROC and identification response analyses showed that, yes, moderate-variation lineups did increase the number of guilty suspect identifications compared to the low-variation lineups, but this also came at a cost: an increase in innocent suspect identifications. Thus, when we consider suspect identifications, our data from Experiment 1 show that increasing the variation in the feature from low to moderate only served to increase subjects'

willingness to identify the suspect, it did not have any discernable impact on ability to discriminate between innocent and guilty suspects.

Why might ability to discriminate between innocent and guilty suspects be similar in the low-variation and moderate-variation lineups? One possibility is that our manipulation was too subtle. The moderate-variation features were rated as less similar to the suspect's feature than the low-variation features, but the faces remained on-screen while subjects made these ratings. Perhaps differences in the variation of the feature are less apparent when people are relying on their memory of a feature that was presented relatively briefly during a mock crime video. Thus, in Experiment 2, we compared do-nothing lineups and low-variation lineups against high-variation lineups, in which we varied the size, appearance and also the location of the feature across the foils.

Finally, the modelling in Experiment 1 illustrated that at least one of our designated innocent suspects was more similar to the culprit than the other foils. This was not problematic for answering our research question and many other studies have used innocent suspects who are relatively more similar to the culprit (e.g., Fitzgerald et al., 2015; Gronlund et al., 2012; Wetmore et al., 2015). The reasoning is that if the police apprehend an innocent suspect, then they are likely to match the lineup foils to the innocent suspect's appearance and this results in a lineup in which the innocent suspect resembles the actual culprit more than the other foils (Clark & Tunnicliff, 2001; Navon, 1992). It did, however, make the identification task difficult. Additional analyses revealed that it was the innocent suspect in the mugging stimulus set who was highly similar to the mugging culprit. After watching the mugging video, 28% of subjects selected our innocent suspect from the low-variation lineup (far higher than the expected 17% for a fair target-absent lineup in which all of the members are equally similar to the culprit).[7] After watching the graffiti video, 13% of subjects selected our innocent suspect from the low-variation lineup (much closer to the expected 17%). Thus, in Experiment 2, we also adjusted the lineups for the mugging video.

[7] Indeed, given that no lineup is *perfectly* fair, the expectation is that a randomly selected designated innocent suspect would, if anything, be chosen less than 17% of the time because the odds are only 1 in 6 that the designated innocent suspect will be the most familiar person in the lineup (Palmer et al., 2013).

Method

Design

We used a 3 (lineup type: low-variation, high-variation, do-nothing) × 2 (target: present, absent) mixed design, with target manipulated within subjects. We recruited as many subjects as possible before the end of the spring term, aiming for at least 400 subjects with usable data in each of the three between-subject conditions.

Subjects

We recruited 1,476 subjects from social network sites who were entered into a prize draw for two £25 Amazon vouchers. All subjects completed the study online. We excluded 68 people (5% in total; between 19–26 subjects in each of the three between-subject conditions) who had experienced technical difficulties while watching the video (n = 16, 1% in total), incorrectly answered an attention check question on the content of the video (n = 11, < 1% in total), or stated that they had seen one of the videos before or had completed the study more than once (n = 41, 3% in total). This resulted in a total sample size of 1,408: between 467 and 472 subjects in each of the three between-subject cells. Table 5.1 shows a demographic breakdown of the sample.

Materials

Videos

We used the same videos as in Experiment 1.

Lineups

For the graffiti culprit, we used the same innocent suspect and foils as in Experiment 1. For the mugging culprit, we adjusted the lineup members. We removed the innocent suspect (top left image in Figure 5.1b) and another foil (top right image in Figure 5.1b) whom we deemed to be very similar-looking to the real culprit (Figure 5.1a). We replaced these with two other foils from the foil pool created in Chapter 2. We then randomly selected one of these six faces to serve as the innocent suspect (see Figure 5.7).

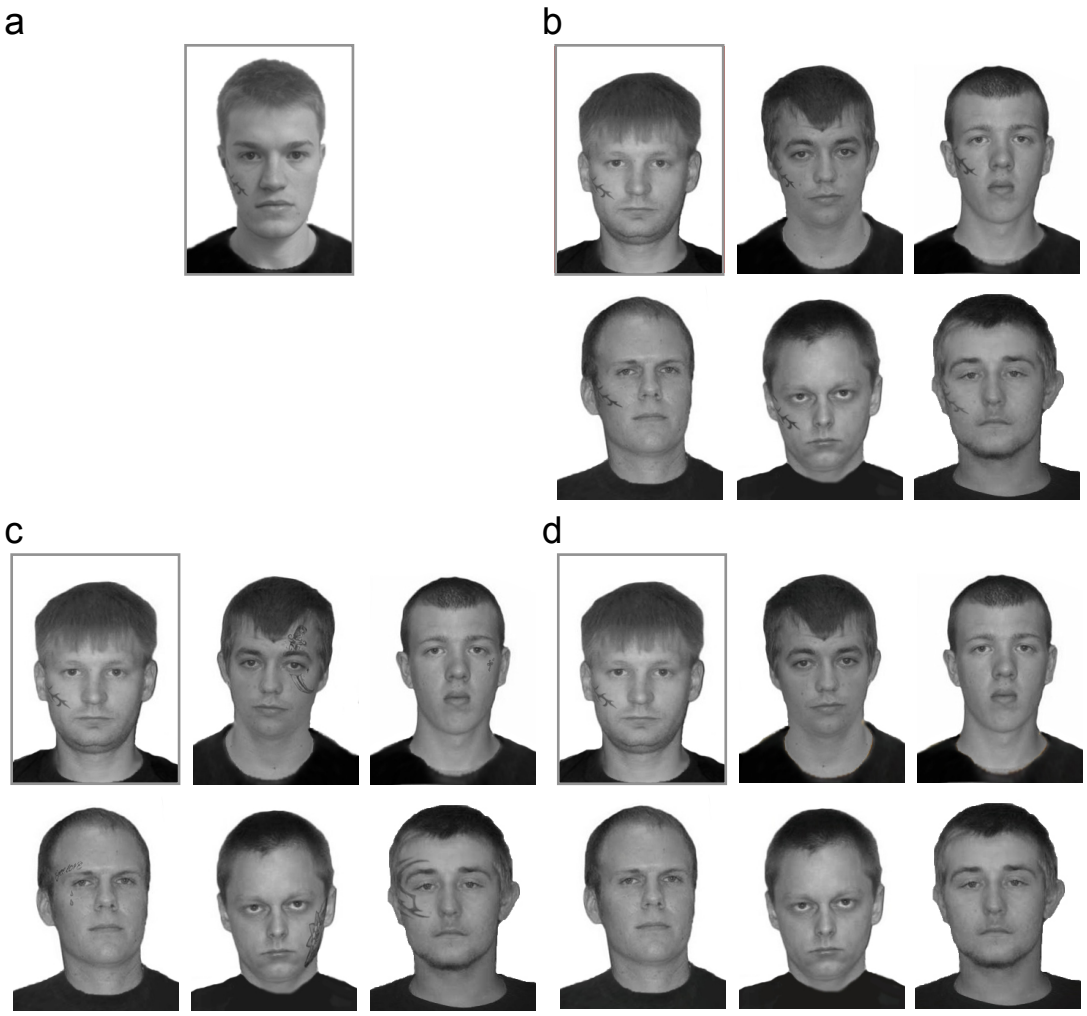

Figure 5.7. (a) A sample culprit, (b) a low-variation lineup, (c) a high-variation lineup, and (d) a do-nothing lineup in Experiment 2. Top left image in each lineup is the innocent suspect with a similar feature to the culprit.

Low-variation and do-nothing lineups were constructed using the same method as in Experiment 1. For high-variation lineups, each foil had a similar feature to the culprit added to them, but each feature was different in size and appearance. For each culprit, we randomly selected five descriptions collected during Experiment 1 that described the feature correctly (i.e., categorised as "code 2"), and used one description for each foil face. Because these descriptions were very vague (e.g., *"Some sort of tattoo on the side of his face…"*), we also searched on the Internet for examples of facial tattoos and black-eyes. We varied the location of the feature across the foils, such that some of the faces had the feature on the right side of their face, the remaining had the feature on the left side of their face. We used this method to ensure that the suspect did not stand out because he was the only lineup member with the feature on the right side of his face.

Lineup fairness. We used the same method as Experiment 1 to confirm that our new mugging lineup members were plausible alternatives to the culprit; Tredoux's *E′* was 3.57, 95% CI [2.85, 4.77] for the target-present lineup (n = 40), and 3.52, 95% CI [2.94, 4.40] for the target-absent lineup (n = 40).

Manipulated images. We used the same method as Experiment 1, but presented subjects with a single target-present lineup (n = 174). Our foil photos did not look doctored; subjects were no better than chance at identifying the non-doctored photo in the low-variation mugging lineup (p = .90), and were significantly worse than chance in the high-variation mugging and graffiti lineups (ps < .04).

Similarity. We used the same method as Experiment 1 (n = 31). Table 5.6 shows that the distinctive features on the low-variation foils were rated as significantly more similar to the suspect's feature than the high-variation foils, in both target-present and target-absent conditions. The distinctive features on each guilty suspect-innocent suspect pair were highly similar (graffiti: M = 8.84, SE = 0.23; mugging: M = *8.55*, SE = 0.32).

Table 5.6
Mean (SE) Ratings of Distinctive Feature Similarity in Experiment 2

		Variation condition			Cohen's *d* and 95% CI		
Suspect	Foil	Low	High	*t*(30)	*d*	LL	UL
Graffiti							
Guilty	1	6.74 (0.39)	2.00 (0.29)	10.18**	2.48	1.68	3.25
	2	8.26 (0.29)	1.58 (0.33)	17.16**	3.82	2.76	4.87
	3	7.90 (0.33)	3.71 (0.35)	7.37**	2.23	1.40	3.03
	4	7.74 (0.26)	1.97 (0.36)	13.53**	3.31	2.34	4.26
	5	7.55 (0.32)	2.52 (0.44)	10.23**	2.36	1.60	3.10
	Average	7.64 (0.21)	2.35 (0.25)	18.19**	4.08	2.96	5.19
Innocent	1	7.23 (0.31)	2.42 (0.35)	10.24**	2.61	1.78	3.43
	2	7.97 (0.30)	1.74 (0.31)	15.23**	3.67	2.62	4.70
	3	7.90 (0.30)	3.55 (0.38)	8.81**	2.29	1.51	3.05
	4	7.74 (0.23)	2.55 (0.41)	12.57**	2.78	1.95	3.59
	5	7.74 (0.27)	1.94 (0.38)	14.92**	3.20	2.28	4.10
	Average	7.72 (0.22)	2.44 (0.26)	18.36**	3.96	2.87	5.03
Mugging							
Guilty	1	8.97 (0.28)	0.68 (0.17)	25.23**	6.37	- [a]	- [a]
	2	8.48 (0.35)	1.42 (0.40)	11.33**	3.42	2.37	4.46
	3	8.42 (0.33)	1.13 (0.31)	18.38**	4.09	2.96	5.20
	4	8.74 (0.28)	1.00 (0.24)	17.70**	5.30	- [a]	- [a]
	5	8.42 (0.23)	2.35 (0.36)	20.19**	3.64	2.65	4.62
	Average	8.61 (0.21)	1.32 (0.22)	23.84**	5.97	- [a]	- [a]
Innocent	1	9.03 (0.23)	1.26 (0.34)	20.37**	4.78	3.48	6.06
	2	8.90 (0.23)	0.94 (0.31)	17.40**	5.24	- [a]	- [a]
	3	8.32 (0.31)	1.19 (0.36)	13.82**	3.81	2.70	4.91
	4	8.81 (0.37)	0.97 (0.26)	15.55**	4.43	3.18	5.67
	5	8.19 (0.37)	2.29 (0.38)	13.94**	2.83	2.01	3.64
	Average	8.65 (0.21)	1.33 (0.26)	21.26**	5.53	- [a]	- [a]

Note. Scale ranged from 0 (*not at all similar*) to 10 (*highly similar*). Cohen's *d* was estimated using the formula: $d = M_{diff} \div s_{av}$, where M_{diff} is the mean difference between the low-variation and high-variation foil, and s_{av} is the average of the standard deviations for the low-variation and high-variation foil (Cumming, 2012). CI = confidence interval; *LL* = lower limit; *UL* = upper limit.
[a] confidence intervals could not be accurately approximated because an unbiased estimate of the population effect size *d* was not between –2 and 2 (Cumming, 2012).
*p = .001, **p < .001.

Procedure

We used the same procedure as Experiment 1.

Results & Discussion

Again, we checked the content of subjects' descriptions, conducted ROC analysis, examined subjects' identification responses and fit a signal detection process model (Wixted & Mickes, 2014). We also examined subjects' ability to judge the accuracy of their suspect identification decisions.

Descriptions

We coded the descriptions using the same method as Experiment 1 (see Table 5.3). There was substantial agreement between the coders, $\kappa = 0.70$ (Landis & Koch, 1977). Again, the majority of the descriptions described the feature correctly including specific location (code 3), but the remaining descriptions were more likely to contain fewer (codes 1 and 2) rather than more (codes 4 and 5) details. Around half of the descriptions provided (46%) described the feature with the level of detail that we used to construct our high-variation lineups (code 2) or provided less detail than we used to construct our high-variation lineups (codes 0 or 1). Again, this shows that the majority of subjects did not describe the distinctive feature in great detail.

ROC analysis

We constructed our ROC curves using the same method as Experiment 1.

Collapsed over the two mock crime scenarios. Figure 5.8 shows that, in the aggregate, the ROC curve for the low-variation lineup lies above the curve for the high-variation lineup, while the do-nothing curve falls below both of these, closer to the chance line. The ROC analysis showed that only low-variation lineups—not high-variation lineups—enhanced subjects' ability to tell the difference between innocent and guilty suspects compared to when nothing was done to prevent the distinctive suspect from standing out. The *p*AUC (specificity = .904) for low-variation lineups (*p*AUC = 0.032, 95% CI: 0.025, 0.039) was significantly greater than the *p*AUCs for both high-variation (*p*AUC = 0.021, 95% CI: 0.015, 0.029, $D = 3.99$, $p < .001$) and do-nothing (*p*AUC = 0.014, 95% CI: 0.009, 0.020, $D = 3.99$, $p < .001$) lineups. The *p*AUCs for high-variation and do-nothing lineups were not significantly different, $D = 1.54$, $p = .12$. Figure 5.8 also shows that subjects were

more willing to identify the suspect in do-nothing lineups compared to low-variation and high-variation lineups, and more willing to identify the suspect in high-variation lineups compared to low-variation lineups. This is because the do-nothing curve extends further right than the low-variation and high-variation curves, and the high-variation curve extends further right than the low-variation curve.

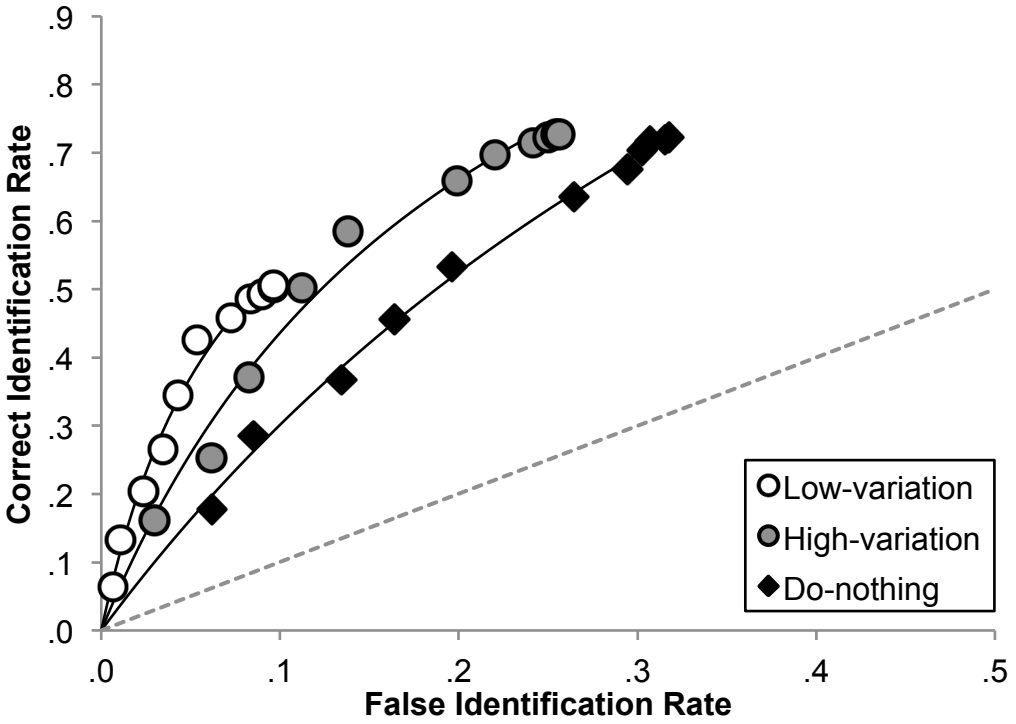

Figure 5.8. Receiver operating characteristic (ROC) curves for the low-variation, high-variation, and do-nothing lineups. Data are collapsed over the mugging and graffiti videos. The dashed line represents chance-level performance.

Separated by mock crime scenario. Figure 5.9 shows a different pattern of results in the (a) mugging and (b) graffiti videos. In the mugging video, only low-variation lineups—not high-variation lineups—enhanced subjects' ability to tell the difference between innocent and guilty suspects compared to when nothing was done to prevent the distinctive suspect from standing out. The *p*AUC (specificity = .931) for low-variation lineups (*p*AUC = 0.027, 95% CI: 0.021, 0.033) was significantly greater than the *p*AUCs for both high-variation (*p*AUC = 0.009, 95% CI: 0.005, 0.014, $D = 4.93$, $p < .001$) and do-nothing (*p*AUC = 0.013, 95% CI: 0.007, 0.019, $D = 3.46$, $p = .001$) lineups. The *p*AUCs for high-variation and do-nothing lineups were not significantly different, $D = 1.01$, $p = .31$. Figure 5.9a also shows that subjects

were more willing to identify the suspect in do-nothing and high-variation lineups compared to low-variation lineups, but, this time, were more willing to identify the suspect in high-variation lineups compared to do-nothing lineups.

For the graffiti video, however, both low-variation and high-variation lineups were equally effective at enhancing subjects' ability to tell the difference between innocent and guilty suspects compared to when nothing was done to prevent the distinctive suspect from standing out. The *p*AUCs (specificity = .877) for both low-variation (*p*AUC = 0.035, 95% CI: 0.023, 0.049, $D = 2.80$, $p = .005$) and high-variation (*p*AUC = 0.048, 95% CI: 0.035, 0.062, $D = 4.09$, $p < .001$) lineups were significantly greater than the *p*AUC for do-nothing lineups (*p*AUC = 0.014, 95% CI: 0.008, 0.023). The *p*AUCs for low-variation and high-variation lineups were not significantly different, $D = 1.38$, $p = .17$. Figure 5.9b also shows that subjects were more willing to identify the suspect in do-nothing lineups compared to low-variation and high-variation lineups, and more willing to identify the suspect in high-variation lineups compared to low-variation lineups.

In sum, the ROC analysis illustrates that, in the aggregate, ability to discriminate between innocent and guilty suspects was better in low-variation lineups compared to high-variation and do-nothing lineups. Descriptively speaking, ability to discriminate between innocent and guilty suspects was better in high-variation lineups than do-nothing lineups, but not significantly so. This pattern of results is predicted by the diagnostic-feature-detection account. Greater variation in the feature across the foils (from low to do-nothing) also increased subjects' willingness to identify the guilty or innocent suspect. However, the pattern of results was different in the mugging and graffiti stimulus sets. In the mugging video, ability to discriminate between innocent and guilty suspects was better in low-variation lineups compared to high-variation and do-nothing lineups. In the graffiti video, ability to discriminate between innocent and guilty suspects was better in both low-variation and high-variation lineups compared to do-nothing lineups. This suggests that the effect of variation in the replicated feature across the foils might depend on the specific encoding and test conditions experienced.

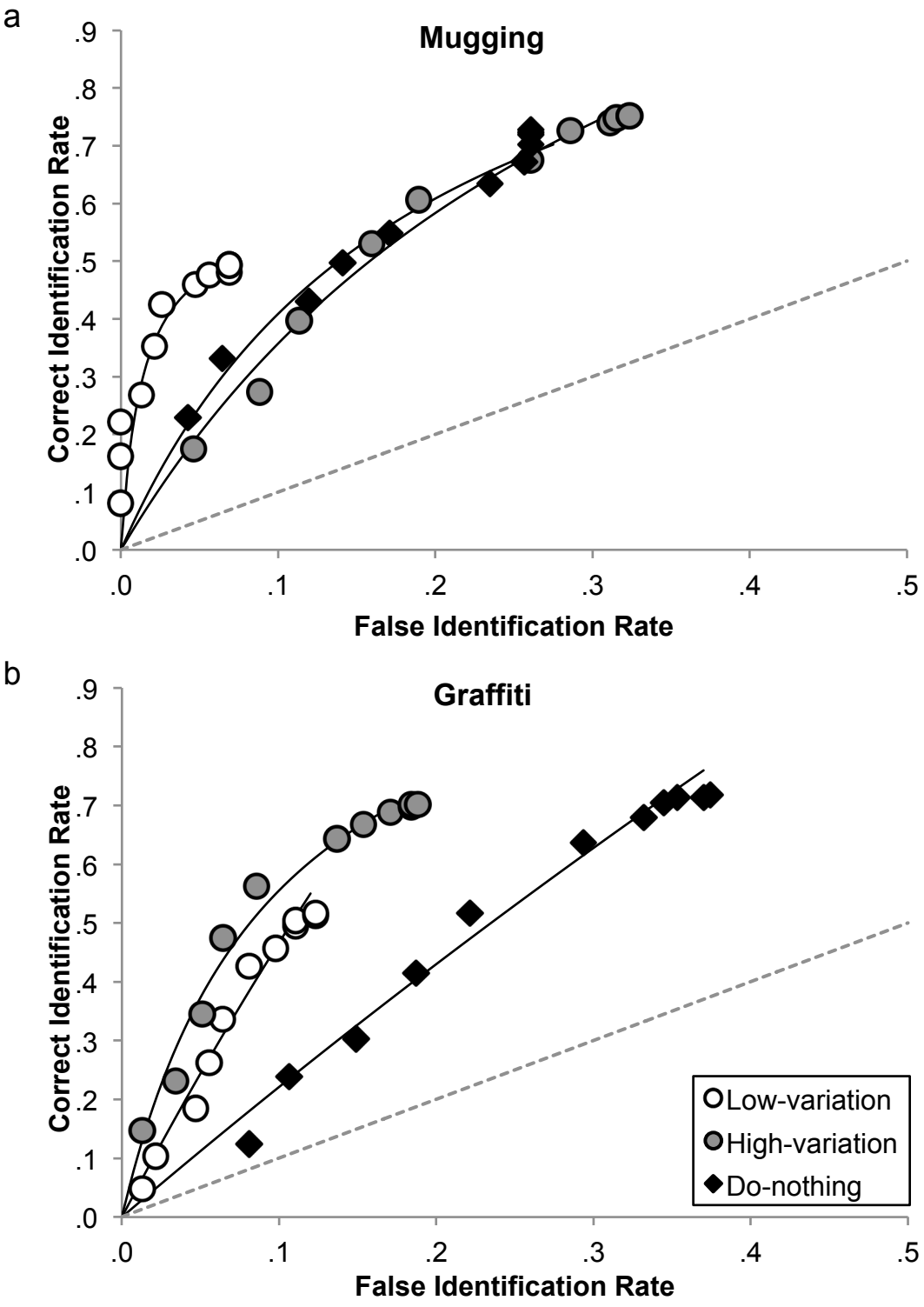

Figure 5.9. Receiver operating characteristic (ROC) curves for the low-variation, high-variation, and do-nothing lineups in the (a) mugging and (b) graffiti videos. The dashed lines represent chance-level performance.

But why did low-variation enhance accuracy more than high-variation in the mugging video, when low-variation and high-variation were equally effective in the graffiti video? One possibility is that the difference between our low- and high-

variation conditions was greater in the mugging lineups than in the graffiti lineups. To check this, we conducted a 2 (video: mugging, graffiti) × 2 (suspect: guilty, innocent) × 2 (condition: low-variation, high-variation) within-subjects ANOVA on the suspect-foil similarity pilot ratings. Ratings were made on an 11-point Likert-type scale that ranged from 0 (*not at all similar*) to 10 (*highly similar*). A main effect of condition indicated that our manipulation was successful; the features on low-variation suspect-foil pairs (M = 8.15, SD = 0.17) were rated as more similar than the features on high-variation pairs, M =1.86, SD = 0.21, F (1, 30) = 546.46, $p < .001$, $\eta_p^2 = .95$. This was, however, qualified by a significant interaction between the video and variation condition, F (1, 30) = 66.80, $p < .001$, $\eta_p^2 = .69$. In the high-variation condition, the bruises on the graffiti suspect-foil pairs were rated as more similar (M = 2.40, SD = 1.37, SE = 0.25) than the tattoos on the mugging suspect-foil pairs, M = 1.32, SD = 1.34, SE = 0.24, t (30) = 4.46, $p < .001$, d = 0.79, 95% CI [0.39, 1.19].[8] In the low-variation condition, the opposite was true: The bruises on the graffiti suspect-foil pairs were rated as less similar (M = 7.68, SD = 1.13, SE = 0.20) than the tattoos on the mugging suspect-foil pairs, M = 8.63, SD = 1.14, SE = 0.21, t (30) = 4.39, $p < .001$, d = 0.84, 95% CI [0.40, 1.26]. In short, there was a larger difference between our low- and high-variation conditions in the mugging lineups than the graffiti lineups and this could explain why low-variation lineups enhanced performance more than high-variation lineups for the mugging scenario, but not the graffiti scenario.

Although the differences between the stimulus sets make an important point about the generalizability of our findings, in this study, we were not interested in examining differences across particular combinations of encoding and test conditions. Rather, we were interested in examining general effects. Therefore, in all subsequent analyses, we collapsed our data over both stimulus sets to examine overall patterns in our data, because this provides the most relevant information about the average impact of the degree of variation in the replicated feature across the foils. Analysing the data in this way is appropriate because in the real world, just like across our stimulus sets, encoding and test conditions are likely to vary widely (Brewer et al., 2010).

---

[8] Again, Cohen's d was estimated using the formula: $d = M_{diff} \div s_{av}$ (see Table 5.6).

Identification responses

We calculated the proportion of suspect identifications, foil identifications and lineup rejections made to the different lineup types. Figure 5.10 shows the identification responses made in the low-variation, high-variation, and do-nothing (a) target-present and (b) target-absent lineups.

Target-present lineups. Figure 5.10a shows that, like in Experiment 1, greater variability in the distinctive feature across the lineup members resulted in more guilty suspect identifications, but, this time, the pattern of identification responses in the high-variation and do-nothing lineups was similar. A 3 (lineup type: low-variation, high-variation, do-nothing) × 3 (identification response: guilty suspect, foil, incorrect rejection) two-way chi-square analysis indicated that lineup technique influenced identification responses, χ^2 (4, N = 1,408) = 82.14, $p < .001$, Contingency Coefficient C = .23. Low-variation lineups resulted in fewer guilty suspect IDs ($z = -3.92$, $p < .001$) and more foil IDs ($z = 5.72$, $p < .001$) and more lineup rejections ($z = 2.32$, $p < .05$), while high-variation and do-nothing lineups led to more guilty suspect IDs (high-variation: $z = 2.01$, $p < .05$; do-nothing: $z = 1.90$, $p > .05$) and fewer foil IDs (high-variation: $z = -3.69$, $p < .001$; do-nothing: $z = -2.01$, $p < .05$) than expected. Specifically, three 2 (lineup type) × 2 (identification response: guilty suspect, foil) two-way chi-square analyses indicated that when subjects made a selection, those who viewed the do-nothing lineup were no more likely to identify the guilty suspect than those who viewed the high-variation lineup, χ^2 (1, N = 763) = 2.00, p = .16, OR = 1.40, 95% CI [0.86, 2.30], but were 3.30 times more likely to identify the guilty suspect than those who viewed the low-variation lineup, χ^2 (1, N = 730) =40.17, $p < .001$, OR = 3.30, 95% CI [2.22, 4.94]. Subjects who viewed the high-variation lineup were 4.61 times more likely to identify the guilty suspect than subjects who viewed the low-variation lineup, χ^2 (1, N = 721) = 56.95, $p < .001$, OR = 4.61, 95% CI [2.99, 7.24].

Target-absent lineups. Figure 5.10b shows that, like in Experiment 1, greater variability in the distinctive feature across the lineup members also resulted in more innocent suspect identifications. A 3 (lineup type: low-variation, high-variation, do-nothing) × 3 (identification response: innocent suspect, foil, correct rejection) two-way chi-square analysis indicated that lineup technique influenced identification responses, χ^2 (4, N = 1,408) = 129.32, $p < .001$, Contingency Coefficient C = .29.

Low-variation lineups resulted in fewer innocent suspect IDs ($z = –5.82$, $p < .001$) and more foil IDs ($z = 6.24$, $p < .001$), while high-variation lineups resulted in fewer foil IDs ($z = –4.93$, $p < .001$) and more lineup rejections ($z = 2.43$, $p < .05$) than expected. Do-nothing lineups resulted in more innocent suspect IDs ($z = 4.30$, $p < .001$) than expected. Specifically, three 2 (lineup type) × 2 (identification response: innocent suspect, foil) two-way chi-square analyses indicated that when subjects made a selection, those who viewed the do-nothing lineup were no more likely to identify the innocent suspect than those who viewed the high-variation lineup, χ^2 (1, $N = 439$) = 1.72, $p = .19$, OR = 0.77, 95% CI [0.51, 1.16], but were 5.87 times more likely to identify the innocent suspect than those who viewed the low-variation lineup, χ^2 (1, $N = 483$) = 77.54, $p < .001$, OR = 5.87, 95% CI [3.83, 9.10]. Subjects who viewed the high-variation lineup were 7.61 times more likely to identify the innocent suspect than subjects who viewed the low-variation lineup, χ^2 (1, $N = 416$) = 88.73, $p < .001$, OR = 7.61, 95% CI [4.80, 12.24].

Echoing the patterns found in Experiment 1, high-variation lineups resulted in the greatest number correct rejections. Three 2 (lineup type) × 2 (identification response: correct, incorrect) two-way chi-square analyses indicated that subjects who viewed the high-variation lineup were 1.49 times more likely to correctly reject the lineup than subjects who viewed the low-variation lineup, χ^2 (1, $N = 939$) = 9.22, $p = .002$, OR = 1.49, 95% CI [1.14, 1.95], and 1.80 times more likely to correctly reject the lineup than subjects who viewed the do-nothing lineup, χ^2 (1, $N = 941$) = 19.98, $p < .001$, OR = 1.80, 95% CI [1.38, 2.35]. Subjects who viewed the low-variation and do-nothing lineups were equally likely to correctly reject the lineup, χ^2 (1, $N = 936$) = 2.06, $p = .15$, OR = 0.83, 95% CI [0.64, 1.08].

In sum, our analyses align with the findings of the ROC analysis and suggest that greater variation in the feature across the foils (from low-variation to do-nothing) increased subjects' willingness to identify the (guilty or innocent) suspect. We also found that low-variation led to more foil identifications in both target-present and target-absent lineups than high-variation and do-nothing lineups, while high-variation led to the greatest number of correct rejections in target-absent lineups.

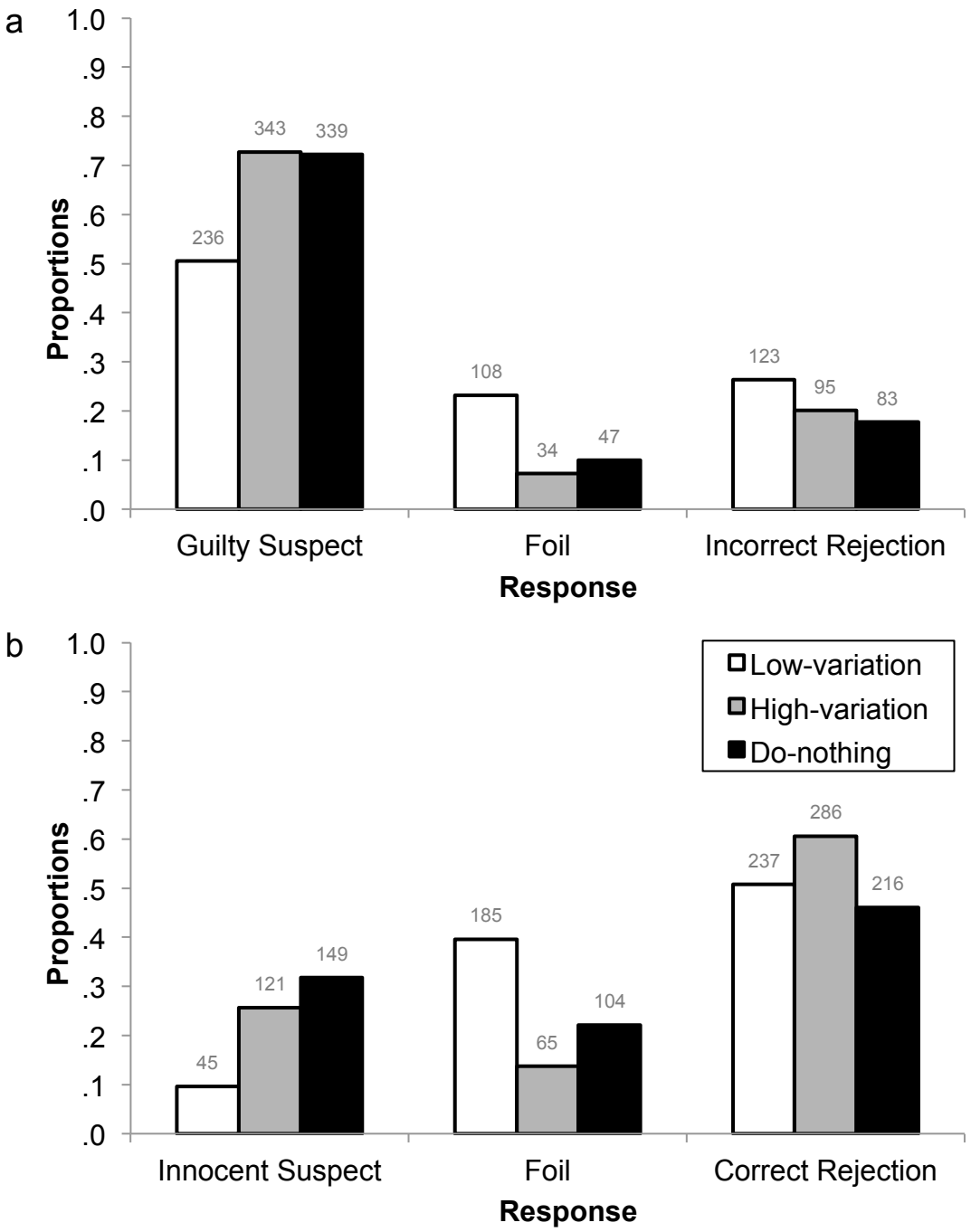

Figure 5.10. Identification responses made in low-variation, high-variation, and do-nothing (a) target-present and (b) target-absent lineups. Data labels are absolute frequencies.

Modelling

To gather more information about subjects' identification performance on the low-variation, high-variation and do-nothing lineups, we followed the same model-fitting procedure outlined in Experiment 1. Our observed data and the values predicted by the best-fitting equal-variance model are shown in Table 5.7, and the

best-fitting parameters and the chi-square goodness-of-fit statistics are shown in Table 5.8. While the simple equal-variance model captured the trends in our data (see Table 5.7), the significant chi-square goodness-of-fit statistics in the left-hand column (full model) of Table 5.8 indicate that our data deviated from the predictions of this simple model, suggesting that a more complex model might fit the data better.[9] Figure 5.11 displays the parameters estimated by the best-fitting equal-variance model for the three lineup types. Before we turn to our main measure of interest—*d′*—it is immediately obvious that the innocent suspect distribution now falls on top of the foil distribution in the low-variation lineups. This indicates that the new innocent suspect in the mugging lineup was equally similar to the culprit as the other foils. That is, the adjustments that we made to the mugging lineup in Experiment 2 were successful.

Now, returning to our main measure of interest—*d′*. It is clear that *d′* declines as the variation in the feature increases (from low to do-nothing). To test whether the observed differences in *d′* were statistically significant, we performed three pairwise comparisons: low-variation versus high-variation, low-variation versus do-nothing, and high-variation versus do-nothing. We fit the same model, allowing the confidence criteria to differ, but constraining *d′* to be equal in the two lineups being compared. The overall χ^2, df and *p* rows in Table 5.8 show the full (unconstrained) and constrained model fit statistics. In comparison to the full model, the constrained model provided a significantly worse fit of the data for the low-variation and high-variation, χ^2 (2) = 148.08, $p < .001$, low-variation and do-nothing, χ^2 (2) = 125.42, $p < .001$, and high-variation and do-nothing, χ^2 (2) = 6.34, $p = .04$, lineup comparisons. These results indicate that increasing the variation in the feature across the foils results in a statistically significant decline in ability to discriminate between innocent and guilty suspects.

It should be noted that the model fit to the high-variation condition—but not the low-variation or do-nothing conditions—was significantly improved by allowing for unequal variance. Again, when we fit an unequal-variance model to our data, we found the same results. In sum, the results of the model fitting are broadly consistent

[9] When we fit the data to the low-variation, high-variation and do-nothing lineup data separately, the model fit the low-variation data well ($p = .18$), but significantly deviated from the observed data in the high-variation ($p = .003$) and do-nothing ($p = .007$) conditions.

with our findings of our ROC analysis when the data were collapsed over the two mock crime scenarios, and suggest that low-variation lineups enhance ability to discriminate between innocent and guilty suspects more than high-variation and do-nothing lineups. However, the model-fitting exercise found that the improvement in discriminability afforded by high-variation lineups compared to do-nothing lineups was statistically significant, while the ROC analysis found that the improvement was not statistically reliable. Nevertheless, our conclusions are the same regardless of whether we use ROC analysis or fit a theoretical model to these data: compared to high-variation and do-nothing lineups, low-variation lineups are the most effective way to enhance subjects' ability to discriminate between innocent and guilty suspects. This pattern of results is predicted by the diagnostic-feature-detection model.

Table 5.7

Observed and Predicted Identification Responses in Each Confidence Bin in the Low-variation, High-variation, and Do-nothing Lineups in Experiment 2

	Low-variation						High-variation						Do-nothing					
	Target present			Target absent			Target present			Target absent			Target present			Target absent		
Confidence	Guilty suspect	Foil	Inc. reject	Innocent suspect	Foil	Correct reject	Guilty suspect	Foil	Inc. reject	Innocent suspect	Foil	Correct reject	Guilty suspect	Foil	Inc. reject	Innocent suspect	Foil	Correct reject
0–20																		
Observed	9.00	17.00	-	6.00	24.00	-	6.00	4.00	-	7.00	8.00	-	9.00	7.00	-	7.00	12.00	-
Predicted	12.61	12.91	-	6.32	25.10	-	8.01	2.81	-	7.98	7.15	-	9.52	4.93	-	10.67	11.19	-
30–40																		
Observed	28.00	23.00	-	14.00	52.00	-	26.00	8.00	-	20.00	18.00	-	32.00	18.00	-	18.00	19.00	-
Predicted	30.56	27.30	-	12.28	47.86	-	25.46	7.67	-	22.15	17.99	-	27.16	12.00	-	26.67	25.06	-
50–60																		
Observed	75.00	35.00	-	9.00	53.00	-	74.00	11.00	-	41.00	22.00	-	84.00	12.00	-	47.00	39.00	-
Predicted	58.81	39.72	-	15.79	59.24	-	66.49	13.56	-	41.94	26.73	-	70.41	21.18	-	51.79	37.57	-
70–80																		
Observed	62.00	22.00	-	11.00	36.00	-	118.00	6.00	-	24.00	8.00	-	80.00	9.00	-	37.00	26.00	-
Predicted	59.09	25.73	-	9.15	32.40	-	92.78	9.31	-	35.04	14.48	-	79.08	12.77	-	38.88	18.57	-
90–100																		
Observed	62.00	11.00	-	5.00	20.00	-	119.00	5.00	-	29.00	9.00	-	134.00	1.00	-	40.00	8.00	-
Predicted	68.11	12.88	-	4.37	14.14	-	137.40	3.36	-	20.79	4.13	-	136.03	6.44	-	32.69	7.76	-
Total																		
Observed	-	-	123.00	-	-	237.00	-	-	95.00	-	-	286.00	-	-	83.00	-	-	216.00
Predicted	-	-	119.28	-	-	240.36	-	-	105.14	-	-	273.61	-	-	89.46	-	-	208.14

Note. The total row displays all reject identification decisions because the model does not account for the confidence level with which lineup rejections are made. Inc. reject = incorrect rejection; Correct reject = correct rejection.

Table 5.8

Full and Constrained (d') Model Fits for the Low-variation vs. High-variation, Low-variation vs. Do-nothing, and High-variation vs. Do-nothing Comparisons in Experiment 2

	Full model		Constrained model	
Estimate	Low-variation	High-variation	Low-variation	High-variation
μ_{guilty} (d')	1.30	1.16	1.10	1.10
μ_{foil}	–0.16	–1.21	–0.72	–0.72
c_1	1.13	0.55	0.68	0.77
c_2	1.24	0.61	0.79	0.82
c_3	1.47	0.78	1.02	0.98
c_4	1.87	1.17	1.43	1.35
c_5	2.35	1.70	1.90	1.86
Overall χ^2	49.28		197.36	
Overall df	26		28	
Overall *p*	.004		< .001	
	Low-variation	Do-nothing	Low-variation	Do-nothing
μ_{guilty} (d')	1.30	0.93	0.96	0.96
μ_{foil}	–0.16	–1.22	–0.79	–0.79
c_1	1.13	0.30	0.60	0.55
c_2	1.24	0.38	0.72	0.62
c_3	1.47	0.58	0.95	0.80
c_4	1.87	1.01	1.36	1.21
c_5	2.35	1.47	1.83	1.64
Overall χ^2	46.12		171.54	
Overall df	26		28	
Overall *p*	.009		<.001	
	High-variation	Do-nothing	High-variation	Do-nothing
μ_{guilty} (d')	1.16	0.93	1.04	1.04
μ_{foil}	–1.21	–1.22	–1.22	–1.22
c_1	0.55	0.30	0.52	0.33
c_2	0.61	0.38	0.58	0.41
c_3	0.78	0.58	0.75	0.62
c_4	1.17	1.01	1.13	1.05
c_5	1.70	1.47	1.66	1.52
Overall χ^2	60.52		66.86	
Overall df	26		28	
Overall *p*	<.001		<.001	

Note. The full model allows d' to differ between the two lineups being compared. The constrained model holds d' constant across the two lineups being compared. Overall χ^2, df and *p* rows represent goodness-of-fit statistics when the model was fit to the two lineups together.

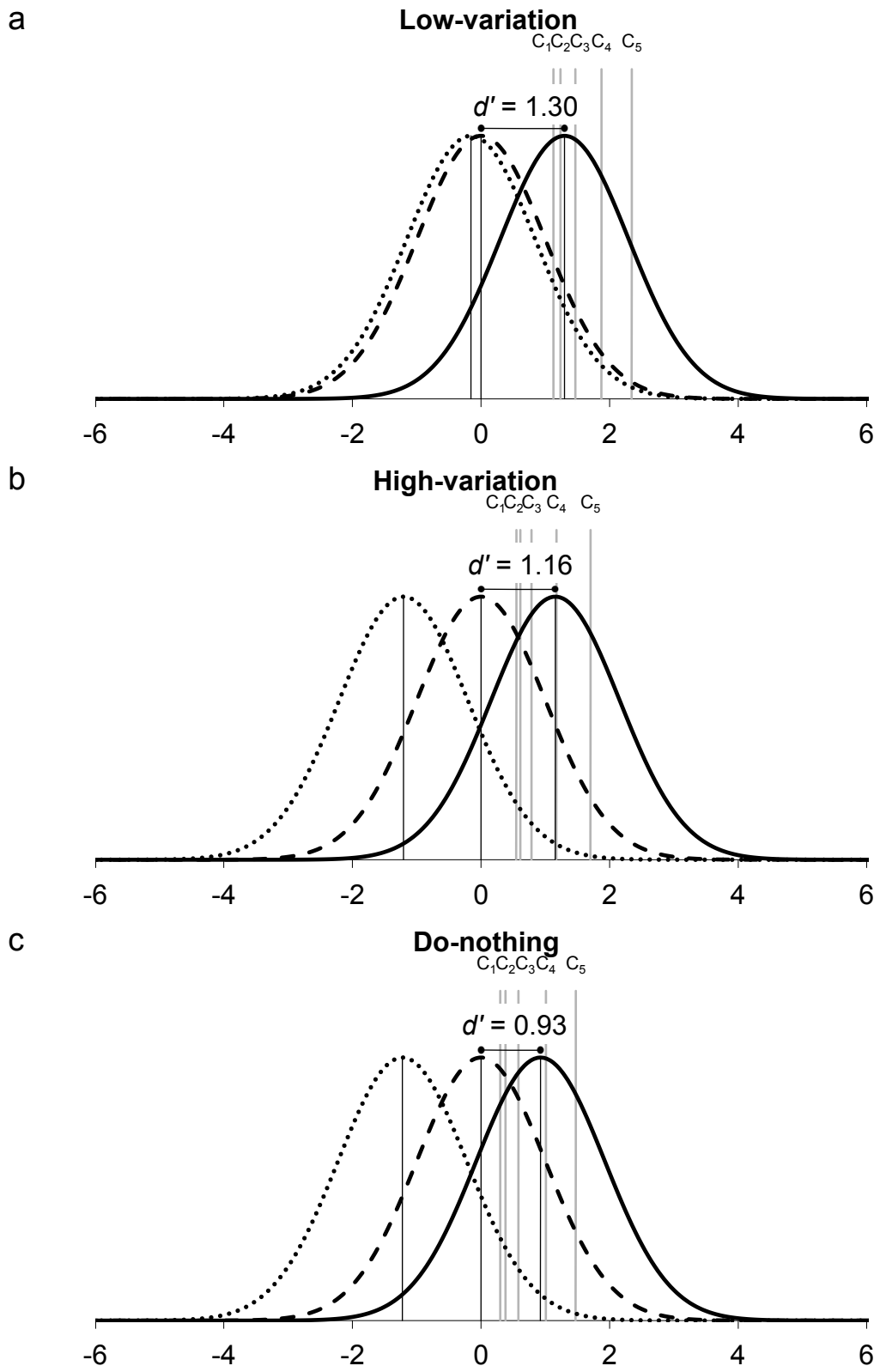

Figure 5.11. Foil, innocent suspect, and guilty suspect distributions for (a) low-variation, (b) high-variation, and (c) do-nothing lineups using the best-fitting equal-variance signal detection model parameters. *d′* measures subjects' ability to discriminate between innocent and guilty suspects. c_1, c_2, c_3, c_4 and c_5 are a set of response criteria that reflect different levels of confidence.

Confidence and accuracy

The ROC analysis and model fitting showed that ability to tell the difference between innocent and guilty suspects was enhanced in low-variation lineups compared to high-variation and do-nothing lineups. According to the diagnostic-feature-detection account, this is because low-variation lineups prevent subjects from relying on the (non-diagnostic) distinctive feature to make their identification decision. If subjects who viewed high-variation and do-nothing lineups fail to lower their confidence judgement despite relying on the (non-diagnostic) feature, then subjects who viewed high-variation and do-nothing lineups will be less accurate at a given level of confidence than subjects who viewed low-variation lineups.

To test this, we constructed confidence-accuracy curves using the same method that we used in Experiment 1. The frequencies of identification responses in each confidence bin are presented in Appendix I. Figure 5.12 shows that low-variation lineups resulted in more accurate suspect identifications than do-nothing lineups at every level of confidence, and more accurate suspect identifications than high-variation lineups at the low (i.e., 0–60% certain) and high (i.e., 90–100% certain) ends of the confidence scale. Generally speaking, this fits with the diagnostic-feature-detection account and suggests that it was not clear that the feature was unhelpful in the high-variation and do-nothing lineups, and therefore subjects failed to make more conservative confidence judgements, despite relying on the feature to make their identification decision.

Interestingly, suspect identifications made with low levels of confidence (i.e., 0–60% certain) from the high-variation lineups were at chance levels of accuracy, such that identifications were less likely to be accurate when they were made from high-variation lineups than do-nothing lineups. One possibility is that subjects who viewed high-variation and do-nothing lineups relied on the feature to make their identification, but, at these low levels of confidence, those in the high-variation gave more liberal confidence judgements for their level of accuracy, compared to those in the do-nothing lineup. Subjects who viewed high-variation lineups may have been relatively more confident in their identification, because these lineups may have given the impression of "fairness". That is, subjects with poor memories viewing high-variation lineups may have been relatively more confident in their decision because they felt like they had searched for and detected a distinctive feature that

matched their memory of the culprit's feature. Conversely, subjects with poor memories viewing do-nothing lineups, may have been relatively less confident in their decision at these low levels of confidence, because the suspect obviously stood out compared to the alternative choices. Further research is required to investigate why high-variation lineups resulted in poorer suspect identification accuracy at low levels of confidence than do-nothing lineups. Nevertheless, we can conclude that low-variation lineups were the most reliable way to enhance the accuracy of suspect identifications across the range of confidence ratings.

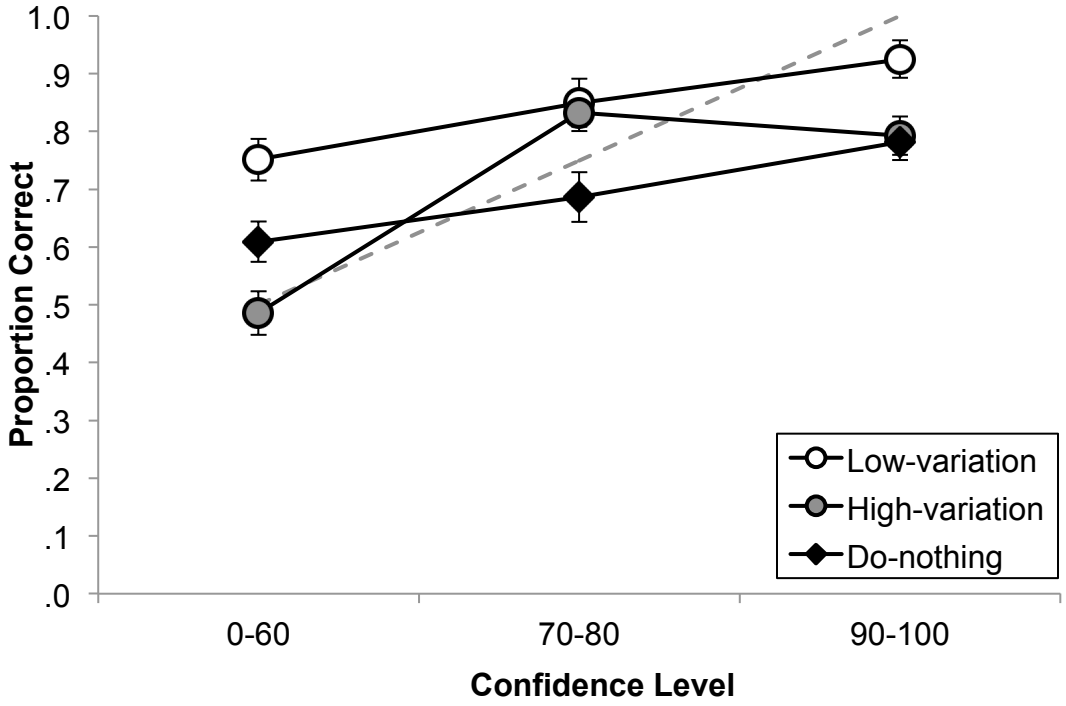

Figure 5.12. Confidence-accuracy curves for suspect identifications in the low-variation, high-variation, and do-nothing lineups. Error bars indicate ±1 SE. The dashed line represents chance accuracy at the lowest confidence bin (i.e., 0–60) and perfect accuracy at the highest confidence bin (i.e., 90–100).

General Discussion

In two experiments we examined how variation in a replicated feature across the foils affected witnesses' identification performance. Compared to doing nothing to prevent a distinctive suspect from standing out, ability to tell the difference between innocent and guilty suspects was enhanced in low-variation and moderate-variation lineups (Experiment 1), but the improvement in discriminability was much less in high-variation lineups (Experiment 2). As such, our research adds to a

growing body of literature which suggests that greater similarity across lineup members can benefit eyewitness identification accuracy (e.g., Clark, 2012; Fitzgerald et al., 2014; Gronlund et al., 2009; Moreland, 2015).

The diagnostic-feature-detection model can explain why little variation in the replicated feature enhances people's ability to tell the difference between innocent and guilty suspects compared to when there is lots of variation in the feature (Wixted & Mickes, 2014). According to the diagnostic-feature-detection account, when the suspect is the only lineup member with the distinctive feature, the feature creates a strong memory-match signal and so people rely on the feature to make their identification decision. This impairs people's ability to tell the difference between innocent and guilty suspects, because the feature is something that both the innocent and guilty suspect share. Conversely, when the feature is replicated with little variation across the foils, it is clear that the feature is not a useful cue to rely upon when making an identification decision because all of the lineup members share that same feature. The distinctive feature is therefore discounted, which enhances people's ability to tell the difference between innocent and guilty suspects compared to doing nothing and leaving the suspect to stand out. When the feature is replicated across the foils with lots of variation, however, the distinctive feature is discounted to a lesser degree because it is less clear that the feature is unhelpful. As such, the improvement in discriminability is less.

Further evidence for the diagnostic-feature-detection account comes from our confidence-accuracy analyses. Generally speaking, greater variation in the feature across the foils resulted in less accurate suspect identifications for a given level of confidence. This fits with a mechanism in which greater variation in the feature makes it less obvious that the feature is an unhelpful cue to use when making an identification decision, because subjects failed to lower their confidence judgements to account for their reduced accuracy. Our results replicate the finding that unfair (do-nothing) lineups distort confidence judgements, because suspect identifications made from do-nothing lineups were less accurate at every level of confidence than those made from low-variation lineups (Chapters 2–4). But our experiments also show that adding a vaguely similar distinctive feature to the foils is not enough to prevent this confidence distortion: only replicating a near identical feature across the

foils was successful at consistently enhancing the accuracy of suspect identifications over the full range of the confidence scale.

Interestingly, though, we did find that high confidence (i.e., 90–100% certain) suspect identifications were equally likely to be accurate in low- and moderate-variation lineups, even though subjects were, overall, more willing to identify suspects in the moderate-variation lineups (Experiment 1). Theoretically, this might suggest that people with stronger memories are less influenced by the degree of variation in the distinctive feature across the members (at least when the variation is only moderate), perhaps because people with stronger memories already seek out more diagnostic facial features on which to base their decision. Further research is required to examine how people search for diagnostic and non-diagnostic cues in lineups and how the strength of the encoded information influences this search. Nevertheless, a key finding is that when subjects were less than 90% certain (i.e., when they were 0–60% and 70–80% certain) their suspect identifications tended to be more accurate when they were made from low-variation lineups compared to moderate-variation lineups.

There were, however, some benefits of increasing the variation in the feature across the lineup members. Compared to low-variation lineups, moderate-variation (Experiment 1) and high-variation (Experiment 2) lineups reduced the number of erroneous foil identifications, and increased the number of correct reject decisions when the real culprit was not present. Many studies have shown that greater similarity across lineup members increases the number of identifications of foils, though the effect of lineup similarity on the number of correct rejections varies across studies (see Fitzgerald et al., 2013, 2015 for reviews). It is likely that the foils in our low-variation lineups felt more familiar (i.e., evoked a greater memory signal) than the foils in the moderate-variation and high-variation lineups, because they each had a feature that was very similar to the culprit's. Indeed, our model fitting supports this account, because there was greater overlap of the memory-strength distributions for foils, innocent suspects and guilty suspects in the low-variation lineups than in the moderate-variation and high-variation lineups. That is, the average distance between the foil and guilty suspect distributions was smaller in the low-variation lineups ($\mu_{guilty} - \mu_{foil} = 1.31$) than the moderate-variation lineups ($\mu_{guilty} - \mu_{foil} = 1.68$) in Experiment 1, and was markedly smaller in the low-variation lineups ($\mu_{guilty} - \mu_{foil}$

= 1.46) than the high-variation lineups ($\mu_{guilty} - \mu_{foil} = 2.37$) in Experiment 2. This indicates that subjects more easily confused foils and suspects when all of the lineup members shared a very similar distinctive feature and can help to explain the greater number of erroneous foil identifications (and the associated decrease in correct reject decisions) in the low-variation lineups.

Many researchers have, however, stressed that erroneously identifying a foil is functionally equivalent to correctly rejecting a target-absent lineup in an applied setting because they both serve to exonerate an innocent suspect (e.g., Fitzgerald et al., 2013, 2015; Wells & Lindsay, 1980; Wells et al., 2006). Arguably, from an applied perspective, research should focus on how lineup procedures influence identifications of suspects, because the Criminal Justice System should use identification procedures that maximise correct identifications of guilty suspects, while minimising incorrect identifications of innocent suspects (Gronlund et al., 2014; National Research Council, 2014). In this regard, the ROC analysis indicated that, compared to low-variation lineups, moderate-variation lineups were equally effective (Experiment 1), but high-variation lineups impaired performance (Experiment 2). In fact, compared to low-variation lineups, moderate-variation lineups (Experiment 1) and high-variation lineups (Experiment 2) increased the number of both guilty and innocent suspect identifications. That is, greater variation in the feature across the foils made subjects more willing to identify the suspect. This is consistent with research in the broader literature that has found that decreasing suspect-foil similarity increases a witness's willingness make an identification decision (Carlson et al., 2008; Clark, 2003, 2012; Clark & Godfrey, 2009; Clark, Rush, & Moreland, 2013; Fitzgerald et al., 2013; Flowe & Ebbesen, 2007). Theoretically, these findings are difficult to reconcile with the replication-with-variation hypothesis, which proposes that varying how the suspect's distinctive feature is replicated across the foils would increase identifications of guilty suspects, without increasing identifications of innocent suspects (Valentine et al., 2009).

What's more, the finding that moderate-variation lineups were equally effective as low-variation lineups, but high-variation lineups impaired people's ability to discriminate between innocent and guilty suspects, highlights an important practical issue: If police officers vary how the feature is replicated across the foils, how do they know how much variation is *too* much? The replication-with-variation

hypothesis suggests that the feature should be replicated within the constraints of a witness's description of the culprit. However, the descriptions we collected from subjects were often vague and generic, despite a prompt that encouraged them to describe the distinctive feature in as much detail as possible. Only a small proportion of the descriptions contained detail about the distinctive feature that was precise enough for us to construct our moderate-variation lineups (8% in Experiment 1; 7% in Experiment 2), whereas almost double the number of descriptions contained the level of detail that we used to construct our high-variation lineups (17% in Experiment 1; 15% in Experiment 2). And it's not just our subjects who were poor at this task. Many other studies have shown that descriptions about facial appearance provided by witnesses can be missing (e.g., R. C. L. Lindsay et al., 1994), vague (e.g., Kuehn, 1974), and inaccurate (e.g., Meissner et al., 2007; van Koppen & Lochun, 1997). If descriptions provided by witnesses of real crimes are similarly poor, then there is likely to be a great deal of variation in how a distinctive feature is replicated over the foils (see Koehnken, Malpass, & Wogalter, 1996 and R. C. L. Lindsay et al., 1994 for similar arguments). Ultimately, our data suggest that a great deal of variation in the feature would impair a witness's ability to tell the difference between who was innocent and who was guilty, compared to lineups in which a near identical feature was added to the foils.

There are, however, two important details to consider when interpreting the results of these two experiments. First, recall that our experiments test a worst-case scenario because the innocent suspects in our target-absent lineups had a very similar distinctive feature to the real culprit. Therefore, while our method ensured the greatest amount of experimental control, it probably overestimates the number of innocent suspect identifications compared to using an innocent suspect who only has a vaguely similar feature to the culprit. Second, recall that we found a different pattern of results in the two stimulus sets in Experiment 2. It is possible that this was because the difference between our low- and high-variation conditions was greater in the mugging lineups than in the graffiti lineups. But it is also likely that the different pattern of results reflect a multitude of complex differences in the encoding and test conditions across the two stimulus sets, such as exposure duration, the similarity of the guilty suspect and the foils, the similarity of the guilty and innocent suspect (Brewer et al., 2010; Brewer & Wells, 2006). Future research should investigate how

varying the similarity of the feature between the culprit and the innocent suspect moderates performance in low-, moderate- and high-variation lineups. Future research could also examine differences across multiple stimulus sets to investigate the upper and lower bounds of accuracy on the different lineup types, or could increase the variability in the test conditions by randomly generating lineups from pools of faces to permit general conclusions to be made. Nevertheless, the critical point is that, in the real world, police officers cannot know whether their suspect is guilty or innocent, nor can they know how similar the feature on the innocent suspect is to the feature on the real culprit, or the specific encoding and test conditions that the witness has encountered. Thus, at the very least, we can conclude that in such worst-case scenarios (a) low-variation lineups do not impair people's ability to discriminate between innocent and guilty suspects and assign appropriate confidence judgements compared to when there is greater variation in the feature across the foils and (b) under some conditions, low-variation lineups can significantly enhance people's ability to discriminate between innocent and guilty suspects and assign appropriate confidence judgements, compared to when there is greater variation in the feature across the foils. Thus, generally speaking, replication-without-variation seems to be the most effective strategy when replicating distinctive features in lineups.

In conclusion, the effect of variation in the replicated feature across the foils might, as with many lineup practices (e.g., Brewer et al., 2010; Brewer & Wells, 2006; Sauer et al., 2008), depend on the particular encoding and test conditions experienced. Nevertheless, our results show that the diagnostic-feature-detection model remains a viable account of eyewitness decision-making, and lend support for the theoretical notion that comparison of similar-looking faces is advantageous in lineups because it allows witnesses to immediately discount many features that are shared by all members (Wixted & Mickes, 2014). But, perhaps most importantly, our study has shed light on the most effective way of replicating a suspect's distinctive feature across the foils. Our data suggest that if you had witnessed a crime committed by a man with a distinctive goatee, then the police could maximise your ability to tell the difference between innocent and guilty suspects and maximise your ability to make an accurate suspect identification at a particular level of confidence, by presenting you with a lineup in which all of the men had the same facial hair.

Chapter 6 :
General Discussion

The aims of this thesis were to investigate how lineup techniques for distinctive suspects influence eyewitness identification performance, and to test the diagnostic-feature-detection model (Wixted & Mickes, 2014). First, a brief summary of the findings from each chapter will prove useful.

Summary

In Chapter 2 we found that all three fair lineup techniques (replication, pixelation, and block) enhanced people's ability to discriminate between innocent and guilty suspects more than unfair (do-nothing) lineups in which nothing was done to prevent the distinctive suspect from standing out. A suspect identification made at a particular level of confidence was more likely to be accurate when it was made from a fair lineup than an unfair lineup.

In Chapter 3 we found that all three fair lineups were equally effective within young, middle-aged and older adults and fair lineups enhanced ability to discriminate between innocent and guilty suspects more than unfair lineups in witnesses of all ages. Again, a suspect identification made at a particular level of confidence was more likely to be accurate when it was made from a fair lineup than an unfair lineup. In fair lineups, the number of erroneous identifications increased with age, and this was due to a decline in discriminability; it was not due an increased willingness to choose. Despite the substantial decline in discriminability with age, older adults made a similar proportion of correct suspect identifications at each level of confidence as young and middle-aged adults on fair lineups.

In Chapter 4 we found that, when the culprit did not have a distinctive feature during the crime, all four lineup techniques (replication, pixelation, block and do-nothing) were equally effective at promoting accurate eyewitness identifications and resulted in a similar proportion of correct suspect identifications at each level of confidence. Performance on each of the fair lineups was similar in all subjects, but subjects who watched a distinctive culprit were less able to discriminate between innocent and guilty suspects in unfair lineups than subjects who watched a non-distinctive culprit. Finally, some subjects who watched a distinctive culprit failed to

recall the distinctive feature but then subsequently relied on it in the identification task when presented with a do-nothing lineup.

Finally, in Chapter 5 we found that, compared to doing nothing to prevent a distinctive suspect from standing out, ability to tell the difference between innocent and guilty suspects was enhanced in low-variation and moderate-variation replication lineups (Experiment 1), but the improvement in discriminability was much less in high-variation replication lineups (Experiment 2). Generally speaking, greater variation in the feature across the foils resulted in less accurate suspect identifications for a given level of confidence.

Practical implications

The studies presented in this thesis provide some of the first empirical tests of how lineup procedures for suspects with distinctive features influence eyewitness identification performance. Therefore, our findings have practical implications for both legal policymakers (e.g., legislators and police chiefs) who decide what type of lineup should be used, and legal decision makers (e.g., judges and jurors) who must determine if identifications are likely to be reliable.

Policymakers

First, our studies highlight the dangers of unfair lineups for distinctive suspects. Current guidelines in many countries and jurisdictions recommend that suspects should not stand out in lineups (e.g., Brooks, 1983; Police and Criminal Evidence Act 1984, Code D, 2011; Technical Working Group for Eyewitness Evidence, 1999). This recommendation was based on much identification research showing that witnesses are more likely to pick the suspect when he looks different to the other lineup members (e.g., Clark, 2012; Doob & Kirshenbaum, 1973; Fitzgerald et al., 2013; Wells, Leippe, & Ostrom, 1979). Indeed, in line with this previous research, in each chapter we found that, when the culprit had a distinctive feature during the crime, doing nothing to prevent a distinctive suspect from standing out made witnesses more likely to identify the suspect, regardless of whether that suspect was innocent or guilty.

More importantly, though, we have shown for the first time that unfair lineups for distinctive suspects can also impair people's ability to tell the difference between innocent and guilty suspects, compared to fair lineup techniques in which

the suspect does not stand out. Many of the procedures that psychological scientists have recommended to the Criminal Justice System—such as ensuring that suspects do not stand out in lineups—have recently been criticised because they reduce witnesses' willingness to make an identification, but they may not improve witnesses' ability to tell the difference between innocent and guilty suspects (see Clark, 2012 and Gronlund et al., 2015 for reviews). This is problematic, because procedures that reduce witnesses' willingness to choose reduce the number of innocent suspect identifications, but they also come at a cost: they reduce the number of guilty suspect identifications. Procedures that improve discriminability, however, are objectively superior because they reduce the number of innocent suspect identifications while also increasing the number of guilty suspect identifications (e.g., Clark, 2012). Therefore, our finding that unfair lineups can also harm people's ability to discriminate, and do not just influence people's willingness to pick the suspect, provides further empirical support for the recommendation that distinctive suspects should not stand out in lineups.

Which method of preventing distinctive suspects from standing out—replication, pixelation, or block—best enhances identification accuracy? In three studies, we found that all three fair lineup techniques led to similar patterns of identification performance (Chapters 2–4). All three fair techniques were equally effective within young, middle-aged and older adults (Chapter 3); when the culprit did not have a feature during the crime; and when the witness failed to describe the culprit's feature (Chapter 4). This suggests that there are multiple ways in which police officers can construct lineups for distinctive suspects in real-world criminal investigations. Current guidelines stipulate that the identification officer overseeing the case has the discretion to choose whether to replicate or conceal a suspect's feature (Police and Criminal Evidence Act 1984, Code D, 2011). Accordingly, some evidence suggests that police officers in the UK more often use concealment (i.e., pixelation or block) techniques, because these are cheaper, faster, require less skill, and can be applied to moving video images, whereas replication techniques cannot (Horry et al., 2013; A. Monaghan, National VIPER User Group, personal communication, August 15, 2016). Given that replication, pixelation and block lineups led to equivalent identification accuracy in our studies, current guidelines

allowing for police officer discretion to select a technique based on practical and financial considerations, appear to be appropriate.

Current guidelines, however, do not specify exactly how police officers should apply their chosen lineup technique, but our research suggests that this may be an important factor in fostering accurate eyewitness identifications. First, we found that when the distinctive feature was replicated with a great deal of variation across the foils, identification performance was not significantly better than doing nothing to prevent the distinctive suspect from standing out (Chapter 5, Experiment 2). Therefore, if a police officer chooses replication, adding a near identical feature to each of the lineup members may be the most effective way to replicate distinctive features in lineups (Chapter 5). Second, previous research found that removing the feature from the face of the suspect led to fewer guilty suspect identifications than replicating the feature over the other lineup members, possibly because the person that subjects believed to be the culprit was missing a distinctive facial feature that they remembered (Badham et al., 2013; Wixted & Mickes, 2014; Zarkadi et al., 2009). Therefore, if a police officer chooses pixelation or block techniques, it follows that the pixelated area or black block should be clearly visible, so that it is obvious to the witness that there could be a feature underneath the concealed area. In brief, further guidance on exactly how a suspect's feature should be replicated, pixelated or block concealed could be a useful addition to current legal guidelines.

Conversely, current guidelines may contain some redundant information. When the witness fails to report a distinctive feature, the Technical Working Group for Eyewitness Evidence (2003) in the US recommends replication, and the Police and Criminal Evidence Act 1984, Code D (2011) in England and Wales recommends concealment (i.e., pixelation or block). Guidelines that promote the use of fair lineups, even when the witness does not describe a distinctive feature, are sensible because police officers can never be certain if the culprit had a distinctive feature at the time of the crime. A witness may have encoded a culprit's distinctive feature but failed to freely recall information about it. And if a witness who encoded the culprit's feature is presented with a lineup in which only the suspect has a distinctive feature, then they will be prone to identifying the suspect, regardless of whether that suspect is innocent or guilty (Chapter 4). Nevertheless, we found that all three fair lineups led to similar identification performance, regardless of whether the culprit

had the feature during the crime and regardless of the content of the subject's description. Therefore, our data suggest that it does not matter whether police officers use replication, pixelation or block techniques in cases when the witness does not include a distinctive feature in their description of the culprit.

In sum, our research suggests that current guidelines for accommodating distinctive suspects in lineups are appropriate. However, it might prove useful to add further information about how best to apply the chosen lineup technique, and to remove rules specifying which particular technique should be used when the witness does not describe a distinctive feature. Notably, current guidelines for accommodating distinctive suspects did not derive from a solid base of scientific evidence about what works best. Therefore, the studies in this thesis provide useful information for policymakers making recommendations for creating lineups for distinctive suspects.

Legal decision makers

Our research also provides useful information for legal decision makers—such as judges and jurors—who are trying to determine if identifications made from lineups for distinctive suspects are likely to be reliable. Studies plotting confidence-accuracy curves show that the confidence judgement taken at the time of the identification decision is often meaningfully related to accuracy (e.g., Horry et al., 2012; Sauerland & Sporer, 2009; Weber & Brewer, 2004; Wixted et al., 2015, 2016; Wixted & Wells, 2016). Even in situations in which memory performance is impaired, people are often aware of this and are able to lower their confidence judgement to account for their poorer accuracy (e.g., Brewer & Wells, 2006; Mickes, 2015, Experiment 1; Palmer et al., 2013; Sauer et al., 2010). But our studies highlight that lineup format can distort confidence judgements (see also Wixted & Wells, 2016). When the culprit had a distinctive feature during the crime, unfair (do-nothing) lineups led to less accurate suspect identifications at almost every confidence level compared to fair lineups (Chapters 2–5). Likewise, replicating a vaguely similar feature across the lineup members (high-variation lineups) led to less accurate suspect identifications at almost every confidence level compared to replicating a near identical feature across the lineup members (low-variation lineups; Chapter 5, Experiment 2). That is, memory performance was impaired on do-nothing and high-variation lineups, but subjects failed to lower their confidence judgement to

account for their poorer accuracy. This illustrates that when witnesses make identifications from do-nothing or high-variation lineups, their confidence judgements might not provide useful information for judges and jurors about the likely accuracy of the identification.

Indeed, one consequence of poorer accuracy at each level of confidence on do-nothing and high-variation lineups, is that subjects made high-confidence suspect identifications when they were not highly likely to be accurate. For instance, subjects' accuracy in the 90–100 confidence bin on do-nothing lineups was only 67% correct in Chapter 2 and only 65% correct for subjects who had watched a distinctive culprit in Chapter 4. Similarly, subjects' accuracy in the 90–100 confidence bin on do-nothing and high-variation lineups in Chapter 5 (Experiment 2) was still only 78% and 81% correct, even though overall identification performance was generally better in Chapter 5 compared to the other chapters. As noted in Chapter 2, this is problematic, because highly confident witnesses can be very influential when judges and jurors make decisions about a suspect's guilt (Brewer & Burke, 2002; Wells, Lindsay, & Ferguson, 1979; see future research section for further discussion about how judges and jurors might interpret eyewitness identification evidence).

Conversely, all three fair lineups—replication, pixelation and block—enabled people to make high-confidence suspect identifications that were highly likely to be accurate (Chapters 2–4). This is useful information for legal decision makers, because it means that high-confidence identifications made on fair lineups for distinctive suspects are likely to be trustworthy. More broadly, then, this finding lends support for the recommendation that police officers should document witness confidence judgements when the initial identification is made, so that judges and jurors can consider this initial confidence judgement when evaluating the identification evidence (e.g., Brewer & Palmer, 2010; D. S. Lindsay et al., 1998; National Research Council, 2014; Wells et al., 1998; Wixted et al., 2015; Wixted & Wells, 2016).

Nevertheless, it is important to note that subjects' accuracy on the fair lineups at the high levels of confidence was (slightly but consistently) lower than is generally reported elsewhere in the literature (see Wixted & Wells, 2016 for a review). On average, our subjects' accuracy in the 90–100 confidence bin on the fair

lineups was 86% correct in Chapter 2; 88% correct in Chapter 3; and was 81% and 93% correct on the low-variation replication lineups in Experiments 1 and 2 in Chapter 5. Wixted and Wells (2016), however, found accuracy was almost always 95% correct or higher. This may suggest that lineups for distinctive suspects are different to standard lineups for non-distinctive suspects in a way that influences the relationship between confidence and accuracy. In our studies, subjects viewed a prominent cue at the time of learning—the distinctive feature—but it was not available to aid their retrieval during the lineup task because the cue was either covered up or appeared on every face. As a result, subjects were using a relatively impoverished retrieval cue because they could not rely on the distinctive feature that they probably encoded. It is possible that such a mechanism could serve to reduce high confidence-accuracy in fair lineups for distinctive suspects (Colloff et al., 2016, supplemental materials).

Alternatively, aspects of our experimental task could have served to reduce high confidence-accuracy in the fair lineups. First, it is possible that our lineup members were more similar looking to the culprit than the lineup members used in previous research. When the lineup members are more similar to the culprit, this could lead to a greater number of high-confidence false identifications, because it is more likely that one lineup member will feel familiar enough to be incorrectly identified with high levels of certainty (Wixted & Wells, 2016). Second, it is possible that making a pre-lineup confidence judgement weakened the relationship between confidence and accuracy on the subsequent lineup task (Bednarz, Carlson, Carlson, Wooten, & Young, 2016). Further research is required to determine whether our somewhat reduced high confidence-accuracy was due to lineups for distinctive suspects being inherently different to lineups for non-distinctive suspects, or was due to a particular aspect of our experimental task.

In sum, although our subjects' accuracy at high levels of confidence on fair lineups was slightly lower than has been observed in the broader literature, we found that identifications made with high levels of confidence on fair lineups for distinctive suspects were still very likely to be accurate. Our data suggest that, when evaluating an identification made at a particular level of confidence, legal decision makers should put more trust in that identification if it was made from a fair (replication, pixelation or block) lineup compared to an unfair (do-nothing) lineup or a high-

variation replication lineup, even if that identification was made with compelling levels of high confidence. As such, it seems that legal decision makers need to be made aware of the lineup technique that was used to gather the eyewitness identification evidence, to know when it is (and it is not) appropriate for them to use a witness's initial confidence judgement as a proxy for their likely accuracy.

Theoretical implications

Because a formal theory of eyewitness discriminability—the diagnostic-feature-detection model—has only been proposed relatively recently, the studies in this thesis provide some of the first empirical tests of predictions made by this model (Wixted & Mickes, 2014). In three studies we found that, when the culprit had the feature during the crime, fair (replication, pixelation, or block) lineup techniques were equally effective and enhanced people's ability to discriminate between innocent and guilty suspects more than unfair do-nothing lineups (Chapters 2–4). We also found that, compared to doing nothing, little variation in how the suspect's feature was replicated across the foils enhanced people's ability to discriminate between innocent and guilty suspects, whereas the improvement was much less when there was greater variation in the feature across the foils (Chapter 5, Experiment 2). These are precisely the patterns of results predicted by the diagnostic-feature-detection model. According to this account, witnesses are better at discriminating between innocent and guilty suspects when they base their decisions on (diagnostic) facial features that differ between innocent and guilty suspects, rather than on (non-diagnostic) facial features that innocent and guilty suspects share. When the suspect is the only person in the lineup with the distinctive feature, the feature creates a strong memory-match signal and so people rely on the feature to make their identification. This impairs their ability to tell the difference between innocent and guilty suspects, because the feature is something that both the innocent and guilty suspect share. By contrast, in the fair lineups, the feature either appears on every lineup member (replication) or none of the lineup members (pixelation or block) and therefore it is clear that the feature non-diagnostic of guilt. The feature is discounted, which promotes reliance on more diagnostic cues and enhances people's ability to discriminate between innocent and guilty suspects. Similarly, when there is greater variation in the replicated distinctive feature, it is less obvious that the feature

is non-diagnostic, which means that the feature is discounted to a lesser degree, and, as such, the improvement in discriminability is less.

Further evidence for the diagnostic-feature-detection account comes from our finding that, when the culprit had the feature during the crime, suspect identifications made from unfair (do-nothing) lineups (Chapters 2–4) and high-variation replication lineups (Chapter 5, Experiment 2) were less likely to be accurate at each level of confidence than identifications made from fair lineups. This fits with a mechanism in which it was unclear that the feature was non-diagnostic in the unfair and high-variation lineups: subjects did not realise that their ability to discriminate between innocent and guilty suspects was impaired on these lineups, so they failed to set a more conservative confidence criteria when making an identification with a particular level of confidence.

Our studies also consolidate the diagnostic-feature-detection account in a number of other ways. First, subjects who watched a non-distinctive culprit in Chapter 4 could not use the distinctive feature in their identification decision, but they made the same pattern of identification responses on replication lineups as those who had watched a distinctive culprit. In accordance with the diagnostic-feature-detection account, this suggests that people who watched a distinctive culprit discounted the distinctive feature when presented with a replication lineup in which all members shared the distinctive feature. Second, the majority of our subjects who watched a distinctive culprit had a memory of the feature because they described it (Chapters 4 and 5). Therefore, the non-significant differences between the replication and concealment (i.e., pixelation and block) techniques are not due to subjects having no memory of the feature.

Moreover, the studies in this thesis also rule out alternative explanations of the results. One alternative mechanism that could explain the poor performance in the unfair lineups observed in Chapters 2 and 3, is that fair lineups guard against the influence of criminal stereotypes, whereas unfair lineups do not. We know that features such as scars, tattoos and pockmarks are deemed to be associated with a stereotypical criminal appearance (MacLin & Herrera, 2006) and criminal appearances can affect sentencing decisions (Funk & Todorov, 2013) and can bias lineup identifications (e.g., Flowe & Humphries, 2011; Flowe et al., 2014). Therefore, when the suspect was the only person in the lineup with a distinctive

feature, it is possible that subjects simply selected the suspect because they deemed him to look most like a criminal. This bias would have been removed in the fair lineups, because all (replication) or none (pixelation, block) of the members had a criminal-looking distinctive feature. Yet, in Chapter 4, subjects who watched a distinctive culprit were significantly worse at discriminating between innocent and guilty suspects on unfair lineups than subjects who watched a non-distinctive culprit. This illustrates that it is ultimately the memory of the culprit's feature that is driving the poor performance observed in our unfair lineups. Thus, non-memory-based explanations, such as unfair lineups permitting subjects to pick the most criminal-looking lineup member, cannot account for our pattern of results.

Another alternative theoretical account—filler siphoning—suggests that fair lineups do not enhance witnesses' underlying ability to discriminate between innocent and guilty suspects. Instead, it suggests that fair lineups yield higher ROC curves than unfair lineups, because the presence of plausible alternatives (the foils, or fillers) in fair lineups siphons some of the incorrect identifications that would otherwise land on the innocent suspect (Wells, 2001; Wells, Smalarz, & Smith, 2015). Filler siphoning was clearly present in our studies, because subjects' incorrect identifications were spread over the foils on the fair lineups, whereas subjects tended to shift those incorrect foil identifications onto the distinctive suspect on the unfair lineups. However, our model fitting (see Appendix A) shows that filler siphoning cannot fully explain our results. The model fitting accounts for the increase in foil identifications in the fair (replication, pixelation and block) lineups compared to the unfair (do-nothing) lineups—that is, it accounts for differences in filler siphoning—yet ability to tell the difference between innocent and guilty suspects was still better in the fair lineups than in the unfair lineups. This illustrates that at least some of the fair lineup advantage is due to enhanced discriminability, as predicted by the diagnostic-feature-detection model (see Wixted & Mickes, 2015 for similar findings, and see future research section for ideas on how to test the diagnostic-feature-detection versus filler siphoning accounts).

Additionally, filler siphoning theory—along with other theories, such as absolute versus relative judgements (Wells, 1984, 1993)—can be conceptualised as a theory of response bias (Wixted & Mickes, 2014). But a theory of discriminability is arguably more useful. Procedures that help improve a witnesses' ability to

discriminate between innocent and guilty suspects, minimise identifications of innocent suspects and maximise identifications of guilty suspects, while procedures that focus on making response bias more conservative, minimise identifications of both innocent suspects and guilty suspects (Clark, 2012). Because policymakers in the Criminal Justice System should seek to employ procedures that enhance discriminability, it is imperative that the field also has a theory of eyewitness discriminability to guide us towards this goal (Gronlund et al., 2015; National Research Council, 2014). Once refined, a theory of eyewitness discriminability could help to improve existing, or develop new identification procedures that enhance the accuracy of eyewitnesses in real criminal investigations (Gronlund et al., 2015; Wixted & Mickes, 2014). Notably, then, the findings in this thesis add to the broader literature and suggest that the diagnostic-feature-detection model remains a viable account of eyewitness discriminability.

Future research

The studies presented in this thesis have important practical and theoretical implications, but several lines of future research would further advance our knowledge about lineups for distinctive suspects and the decision-making processes of eyewitnesses. First, we replicated the finding that fair (replication, pixelation and block) lineups enhanced performance more than unfair (do-nothing) lineups, in four chapters. Nevertheless, it is important to consider that this finding was replicated using a relatively restricted range of encoding and test conditions compared to the variety of encoding and test conditions that could be experienced in real life criminal events. We collapsed our data over multiple stimulus sets to examine general trends in performance, but future research could also compare performance across different stimulus sets to help us to examine the upper and lower bounds of identification performance on lineups for distinctive suspects (e.g., see Brewer et al., 2010). For instance, we know that certain features, such as the hair and face outline, are important for accurate recognition of unfamiliar faces, and so concealing large portions of these regions might impair eyewitness performance compared to replicating distinctive features in these regions (see Johnston & Edmonds, 2009 for a review). Therefore, future research could examine whether the effectiveness of replication, pixelation and block techniques depends on the size and location of the suspect's distinctive feature. We also know that the similarity of lineup members can

influence eyewitness decision-making (see Fitzgerald et al., 2013, 2015 for a meta-analysis and review), but in Chapter 5 we used pre-designated lineup members and our innocent suspects each had a very similar distinctive feature to the culprit. Therefore, future research could examine whether the efficacy of replication-with-variation depends on the similarity of the lineup members and the similarity of the distinctive feature that is shared by the culprit and the innocent suspect. In short, a full understanding of how lineups for distinctive suspects may influence eyewitness identifications in real life criminal investigations requires systematic investigation of the different lineup techniques across a range of forensically relevant variables.

Furthermore, the studies presented here provide information about the most effective methods of digitally altering lineup images for distinctive suspects; what remains unclear, is how best to conduct the lineup itself. In England and Wales, for instance, witnesses are told when the lineup images have been digitally modified (A. Monaghan, National VIPER User Group, personal communication, August 15, 2016; C. Wilkinson, Northamptonshire police, personal communication, August 17, 2016). But those who have witnessed a distinctive culprit might interpret these instructions to mean that the lineup images have been digitally modified because the distinctive culprit that they described to the police is in the lineup. Research shows that seemingly minor changes to lineup instructions can make witnesses more or less likely to believe that the culprit is in the lineup and can influence eyewitness identification performance (Goshen-Gottstein & Groner, 2016; Malpass & Devine, 1981; Steblay, 1997). Moreover, the Police and Criminal Evidence Act 1984, Code D (2011) also permits witnesses to view one lineup image without digital modification. But additional recognition tests may not benefit eyewitness identification accuracy. For instance, one study showed that when subjects were required to view a sequential lineup twice compared to once, subjects were more likely to make a positive identification, but were not more likely to be accurate. Moreover, subjects who chose to view a sequential lineup a second time were less able to discriminate between the culprit and the foils than subjects who chose to view the lineup only once (Horry, Brewer, Weber, & Palmer, 2015). Given that the way in which a lineup is conducted can influence eyewitness identification performance, research is needed to examine whether the current instructions and

procedures for lineups for distinctive suspects, benefit, or indeed harm, eyewitness identification accuracy.

It is also not currently clear how judges and jurors interpret eyewitness identifications made from lineups for distinctive suspects. For instance, we noted that the number of high-confidence errors made on do-nothing and high-variation replication lineups was problematic, because identifications made with high levels of confidence are likely to be very influential on legal decision makers (e.g., Brewer & Burke, 2002; Wells, Lindsay, & Ferguson, 1979). Recent research also suggests that interpretations of confidence judgements made by witnesses can be influenced by the context in which the confidence judgement is made (Cash & Lane, 2016; Dodson & Dobolyi, 2015). It is possible, for example, that judges and jurors might recognise an unfair (do-nothing) lineup and adjust their evaluation of the quality of the eyewitness identification evidence to account for the inherent suggestiveness of an unfair lineup procedure (Devenport, Stinson, Cutler, & Kravitz, 2002). Yet high-variation replication lineups, unlike do-nothing lineups, might give the impression of fairness because all of the members have a distinctive feature. As such, it is possible that legal decision makers may attach more weight to a high-confidence identification made from a high-variation lineup than from a do-nothing lineup, even though suspect identifications were similarly poor on both lineup types. In addition, we also noted that all three fair (replication, pixelation and block) lineups provided reliable information for judges and jurors, because all three techniques helped subjects to make highly confident identifications when they were very likely to be accurate. But, currently, we do not know if legal decision makers attach the same weight to high-confidence identifications made from replication, pixelation and block lineups. Given that eyewitness identification evidence heavily influences verdict decisions (e.g., Devlin, 1976; Pozzulo et al., 2006, 2009), future research should examine how jurors interpret confidence judgements made by witnesses who have viewed different types of lineups for distinctive suspects.

Finally, additional research is required to further our theoretical understanding of diagnostic-feature-detection and filler siphoning processes in eyewitness decision-making. Our modelling demonstrated that at least some of the fair lineup advantage was likely due to diagnostic-feature-detection. Yet, filler siphoning was present in our data and it is possible (and indeed likely) that filler siphoning plays a role in

eyewitness decision-making in the real world (e.g., Smith, Wells, Lindsay, & Penrod, 2016; Wells, Smalarz, & Smith, 2015). Future research could experimentally manipulate factors to empirically test these two theories. For example, subjects could be presented with either a single image of the suspect, or an image of the suspect surrounded by a number of similar-looking faces, before being asked to make a yes/no decision about whether the suspect is the culprit. The diagnostic-feature-detection account proposes that comparison of facial features across lineup members enhances people's ability to discriminate between innocent and guilty suspects. Thus, the diagnostic-feature-detection model would predict that simply presenting similar-looking faces around the suspect would enhance discriminability, even when subjects cannot identify one of the other faces. The filler siphoning account, however, proposes that the opportunity to identify other plausible alternatives (i.e., the foils) in fair lineups reduces the number incorrect identifications that would otherwise land on the innocent suspect. Thus, the filler siphoning account does not predict any advantage of presenting similar-looking faces around the suspect when there is no opportunity for erroneous identifications to land on the other faces. Put simply, if presenting similar-looking faces alongside the suspect enhances people's ability to discriminate between innocent and guilty suspects compared to presenting the suspect alone, then this would provide evidence for the diagnostic-feature detection account, not the filler siphoning account. Given the importance of theory development in guiding practical recommendations (Gronlund et al., 2015), research should continue to examine the contribution of diagnostic-feature-detection and filler siphoning processes in eyewitness identification tasks.

Concluding remarks

The aims of this thesis were to investigate how lineup techniques for distinctive suspects influence eyewitness identification performance, and to test the diagnostic-feature-detection model (Wixted & Mickes, 2014). The results of our studies converge to suggest that all three fair lineup techniques currently used by the police to accommodate distinctive suspects—replication, pixelation and block—are equally effective and, when the culprit has the feature at the time of the crime, all enhance people's ability to discriminate between innocent and guilty suspects more than doing nothing to prevent a distinctive suspect from standing out. All three fair lineup techniques also enable people to make highly confident decisions when they

are likely to be accurate. Our findings align with the predictions of the diagnostic-feature-detection model which suggests that comparison of facial features across lineup members benefits identification performance, because it allows witnesses to see, and then discount, the non-diagnostic facial features that are shared by all members (Wixted & Mickes, 2014). Put succinctly, both practically and theoretically, it does not seem to matter if police officers replicate, pixelate, or block conceal features in lineups, but they must prevent the distinctive suspect from standing out.

Chapter 7 :
References

Acierno, R., Hernandez, M. A., Amstadter, A. B., Resnick, H. S., Steve, K., Muzzy, W., & Kilpatrick, D. G. (2010). Prevalence and correlates of emotional, physical, sexual, and financial abuse and potential neglect in the United States: The national elder mistreatment study. *American Journal of Public Health*, *100*, 292–297. doi:10.2105/AJPH.2009.163089

Adams-Price, C. (1992). Eyewitness memory and aging: Predictions of accuracy in recall and person recognition. *Psychology and Aging*, *7*, 602–608. doi:10.1037//0882-7974.7.4.602

Badham, S. P., Wade, K. A., Watts, H. J. E., Woods, N. G., & Maylor, E. A. (2013). Replicating distinctive facial features in lineups: Identification performance in young versus older adults. *Psychonomic Bulletin & Review*, *20*, 289–295. doi:10.3758/s13423-012-0339-2

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism Spectrum Quotient (AQ): Evidence from Asperger syndrome/high functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, *31*, 5–17. doi:10.1023/A:1005653411471

Bartlett, J. C. (2014). The older eyewitness. In T. J. Perfect & D. S. Lindsay (Eds.), *The SAGE handbook of applied memory* (pp. 654–674). doi:10.4135/9781446294703.n36

Bartlett, J. C., & Fulton, A. (1991). Familiarity and recognition of faces in old age. *Memory & Cognition*, *19*, 229–238. doi:10.3758/BF03211147

Bartlett, J. C., Hurry, S., & Thorley, W. (1984). Typicality and familiarity of faces. *Memory & Cognition, 12,* 219–228. doi:10.3758/BF03197669

Bartlett, J. C., & Memon, A. (2007). Eyewitness memory in young and older eyewitnesses. In R. C. L. Lindsay., D. F. Ross., J. D. Read, & M. P. Toglia (Eds.), *Handbook of eyewitness psychology: Memory for people* (Vol. 2, pp. 309–338). Mahwah, NJ: Erlbaum Inc.

Bartlett, J. C., Strater, L., & Fulton, A. (1991). False recency and false fame of faces in young adulthood and old age. *Memory & Cognition*, *19*, 177–188. doi:10.3758/BF03197115

Bednarz, J., Carlson, C., Carlson, M., Wooten, A., & Young, D. (2016, March). *Eyewitness confidence and accuracy: An evaluation of pre- versus post-lineup confidence.* Poster presented at the meeting of American Psychology-Law Society, Atlanta, GA.

Bothwell, R. K., Deffenbacher, K. A., & Brigham, J. C. (1987). Correlation of eyewitness accuracy and confidence: Optimality hypothesis revisited. *Journal of Applied Psychology, 72,* 691–695. doi:10.1037/0021-9010 .72.4.691

Boutet, I., Taler, V., & Collin, C. A. (2015). On the particular vulnerability of face recognition to aging: A review of three hypotheses. *Frontiers in Psychology*, *6*, 1139. doi:10.3389/fpsyg.2015.01139 Article:1139.

Brewer, N., & Burke, A. (2002). Effects of testimonial inconsistencies and eyewitness confidence on mock-juror judgments. *Law and Human Behavior*, *26*, 353–364. doi:10.1023/A:1015380522722

Brewer, N., Keast, A., & Rishworth, A. (2002). The confidence-accuracy relationship in eyewitness identification: The effects of reflection and disconfirmation on correlation and calibration. *Journal of Experimental Psychology: Applied*, *8*, 44–56. doi:10.1037/1076-898X.8.1.44

Brewer, N., Keast, A., & Sauer, J. D. (2010). Children's eyewitness identification performance: Effects of a *Not Sure* response option and accuracy motivation. *Legal and Criminological Psychology*, *15*, 261–277. doi:10.1348/135532509X474822

Brewer, N., & Palmer, M. A. (2010). Eyewitness identification tests. *Legal and Criminological Psychology*, *15*, 77–96. doi:10.1348/135532509X414765

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied*, *12*, 11–30. doi:10.1037/1076-898X.12.1.11

Brigham, J. C., Ready, D. J., & Spier, S. A. (1990). Standards for evaluating the fairness of photograph lineups. *Basic and Applied Social Psychology, 11,* 149–163. doi:10.1207/s15324834basp1102_3

Brooks, N. (1983). *Police guidelines: Pretrial eyewitness identification procedures.* Ottawa, Canada: Law Reform Commission of Canada. Retrieved from http://www.lareau-legal.ca/Pretrial.pdf

Cabinet Office. (2015). *What works network.* Retrieved from https://www.gov.uk/guidance/what-works-network#the-what-works-network

Carlson, C. A., Gronlund, S. D., & Clark, S. E. (2008). Lineup composition, suspect position, and the sequential lineup advantage. *Journal of Experimental Psychology: Applied*, *14*, 118–128. doi:10.1037/1076-898X.14.2.118

Cash, D. K., & Lane, S. M. (2016). Context influences interpretation of eyewitness confidence statements. *Law and Human Behavior*. Advance online publication. doi:10.1037/lhb0000216

Chandler, C. C. (1994). Studying related pictures can reduce accuracy, but increase confidence, in a modified recognition test. *Memory & Cognition*, *22*, 273–280. doi:10.3758/BF03200854

Charman, S. D., & Wells, G. L. (2007). Eyewitness lineups: Is the appearance-change instruction a good idea? *Law and Human Behavior*, *31*, 3–22. doi:10.1007/s10979-006-9006-3

Charman, S. D., Wells, G. L., & Joy, S. W. (2011). The dud effect: Adding highly dissimilar fillers increases confidence in lineup identifications. *Law and Human Behavior*, *35*, 479–500. doi:10.1007/s10979-010-9261-1

Clark, S. E. (2003). A memory and decision model for eyewitness identification. *Applied Cognitive Psychology, 17,* 629–654. doi:10.1002/acp.891

Clark, S. E. (2005). A re-examination of the effects of biased lineup instructions in eyewitness identification. *Law and Human Behavior*, *29*, 575–604. doi:10.1007/s10979-005-7121-1

Clark, S. E. (2008). The importance (necessity) of computational modeling for eyewitness identification research. *Applied Cognitive Psychology, 22,* 803–813. doi:10.1002/acp.1484

Clark, S. E. (2012). Costs and benefits in eyewitness identification reform: Psychological science and public policy. *Perspectives on Psychological Science, 7,* 238–259. doi:10.1177/174569161243958

Clark, S. E., Erickson, M. A., & Breneman, J. (2011). Probative value of absolute and relative judgments in eyewitness identification. *Law and Human Behavior*, *35*, 364–380. doi:10.1007/s10979-010-9245-1

Clark, S. E., & Godfrey, R. D. (2009). Eyewitness identification evidence and innocence risk. *Psychonomic Bulletin & Review*, *16*, 22–42. doi:10.3758/PBR.16.1.22

Clark, S. E., Moreland, M. B., & Gronlund, S. D. (2014). Evolution of the empirical and theoretical foundations of eyewitness identification reform. *Psychonomic Bulletin & Review, 21,* 251–267. doi:10.3758/s13423-013-0516-y

Clark, S. E., Rush, R. A., & Moreland, M. B. (2013). Constructing the lineup: Law, reform, theory, and data. In B. L. Cutler (Ed.), *Reform of eyewitness identification procedures* (pp. 87–112). Washington, D.C.: American Psychological Association.

Clark, S. E., & Tunnicliff, J. L. (2001). Selecting lineup foils in eyewitness identification experiments: Experimental control and real-world simulation. *Law and Human Behavior*, *25*, 199–216. doi:10.1023/A:1010753809988

Cochran, W. G. (1952). The χ^2 test of goodness of fit. *Annals of Mathematical Statistics, 25,* 315–345.

Colloff, M. F., Wade, K. A., & Strange, D. (2016). Unfair lineups make witnesses more likely to confuse innocent and guilty suspects. [Supplemental material]. *Psychological Science, 27,* 1227–1239. doi:10.1177/0956797616655789

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* New York: Routledge.

Cutler, B. L., & Penrod, S. D. (1995). *Mistaken identification: The eyewitness, psychology, and the law.* New York: Cambridge University Press.

Davies, G., & Flin, R. (1984). The man behind the mask: Disguise and face recognition. *Human Learning: Journal of Practical Research & Applications*, *3*, 83–95.

Davis, J. P., Valentine, T., Memon, A., & Roberts, A. J. (2015). Identification on the street: A field comparison of police street identifications and video line-ups in England. *Psychology, Crime & Law*, *21*, 9–27. doi:10.1080/1068316X.2014.915322

Deffenbacher, K. A., Bornstein, B. H., Penrod, S. D., & McGorty, E. K. (2004). A meta-analytic review of the effects of high stress on eyewitness memory. *Law and Human Behavior*, *28*, 687–706. doi:10.1007/s10979-004-0565-x

Devenport, J. L., Stinson, V., Cutler, B. L., & Kravitz, D. A. (2002). How effective are the cross-examination and expert testimony safeguards? Jurors' perceptions of the suggestiveness and fairness of biased lineup procedures. *Journal of Applied Psychology*, *87*, 1042–1054. doi:10.1037/0021-9010.87.6.1042

Devlin, P. (1976). *Report to the secretary of state for the home department of the departmental committee on evidence of identification in criminal cases.* London, England: Her Majesty's Stationery Office.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General, 115,* 107–117. doi:10.1037/0096-3445.115.2.107

Dobolyi, D. G., & Dodson, C. S. (2013). Eyewitness confidence in simultaneous and sequential lineups: A criterion shift account for sequential mistaken identification overconfidence. *Journal of Experimental Psychology: Applied*, *19*, 345–357. doi:10.1037/a0034596

Dodson, C. S., Bawa, S., & Krueger, L. E. (2007). Aging, metamemory, and high-confidence errors: A misrecollection account. *Psychology and Aging, 22,* 122–133. doi:10.1037/0882-7974.22.1.122

Dodson, C. S., Bawa, S., & Slotnick, S. D. (2007). Aging, source memory, and misrecollections. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 169–181. doi:10.1037/0278- 7393.33.1.169

Dodson, C. S., & Dobolyi, D. G. (2015). Misinterpreting eyewitness expressions of confidence: The featural justification effect. *Law and Human Behavior, 39,* 266–280. doi:10.1037/lhb0000120

Dodson, C. S., & Krueger, L. E. (2006). I misremember it well: Why older adults are unreliable eyewitnesses. *Psychonomic Bulletin & Review*, *13*, 770–775. doi:10.3758/BF03193995

Doob, A. N., & Kirshenbaum, H. M. (1973). Bias in police lineups—partial remembering. *Journal of Police Science and Administration*, *1*, 287–293.

Dywan, J., & Jacoby, L. (1990). Effects of aging on source monitoring: Differences in susceptibility to false fame. *Psychology and Aging*, *5*, 379–387. doi:10.1037/0882-7974.5.3.379

Ebbesen, E. B., & Flowe, H. D. (2002). *Simultaneous v. sequential lineups: What do we really know?* Retrieved from https://dspace.lboro.ac.uk/2134/20167

Edmonds, E. C., Glisky, E. L., Bartlett, J. C., & Rapcsak, S. Z. (2012). Cognitive mechanisms of false facial recognition in older adults. *Psychology and Aging*, *27*, 54–60. doi:10.1037/a0024582

Ellis, H. D. (1975). Recognizing faces. *British Journal of Psychology, 66,* 409–426. doi:10.1111/j.2044-8295.1975.tb01477.x

Erickson, W. B., Lampinen, J. M., & Moore, K. N. (2015). Eyewitness identifications by older and younger adults: A meta-analysis and discussion. *Journal of Police and Criminal Psychology, 31*, 108–121. doi:10.1007/s11896-015-9176-3

Fawcett, J. M., Russell, E. J., Peace, K. A., & Christie, J. (2013). Of guns and geese: A meta-analytic review of the 'weapon focus' literature. *Psychology, Crime & Law*, *19*, 35–66. doi:10.1080/1068316X.2011.599325

Field, A. P., Miles, J. N. V., & Field, Z. C. (2012). *Discovering statistics using R: and sex and drugs and rock 'n' roll.* London, England: Sage publications.

Fife, D., Perry, C., & Gronlund, S. D. (2014). Revisiting absolute and relative judgments in the WITNESS model. *Psychonomic Bulletin & Review*, *21*, 479–487. doi:10.3758/s13423-013-0493-1

Fitzgerald, R. J., Oriet, C., & Price, H. L. (2015). Suspect filler similarity in eyewitness lineups: A literature review and a novel methodology. *Law and Human Behavior*, *39*, 62–74. doi:10.1037/lhb0000095

Fitzgerald, R. J., & Price, H. L. (2015). Eyewitness identification across the life span: A meta-analysis of age differences. *Psychological Bulletin*, *141*, 1228–1265. doi:10.1037/bul0000013

Fitzgerald, R. J., Price, H. L., Oriet, C., & Charman, S. D. (2013). The effect of suspect filler similarity on eyewitness identification decisions: A meta-analysis. *Psychology, Public Policy, and Law*, *19*, 151–164. doi:10.1037/a0030618

Fitzgerald, R. J., Whiting, B. F., Therrien, N. M., & Price, H. L. (2014). Lineup member similarity effects on children's eyewitness identification. *Applied Cognitive Psychology, 28,* 409–418. doi:10.1002/acp.3012

Flowe, H. D., & Ebbesen, E. B. (2007). The effect of lineup member similarity on recognition accuracy in simultaneous and sequential lineups. *Law and Human Behavior*, *31*, 33–52. doi:10.1007/s10979-006-9045-9

Flowe, H. D., Ebbesen, E. B., Libuser, M., Burke, C., & VanNess, N. (2010). *Testing the reflection assumption: A comparison of eyewitness ecology in the laboratory and the field.* Retrieved from https://dspace.lboro.ac.uk/2134/20195

Flowe, H. D., & Humphries, J. E. (2011). An examination of criminal face bias in a random sample of police lineups. *Applied Cognitive Psychology, 25,* 265–273. doi:10.1002/acp.1673

Flowe, H. D., Klatt, T., & Colloff, M. F. (2014). Selecting fillers on emotional appearance improves lineup identification accuracy. *Law and Human Behavior, 38,* 509–519. doi:10.1037/lhb0000101

Flowe, H. D., Mehta, A., & Ebbesen, E. B. (2011). The role of eyewitness identification evidence in felony case dispositions. *Psychology, Public Policy, and Law*, *17*, 140–159. doi:10.1037/a0021311

Fulton, A., & Bartlett, J. C. (1991). Young and old faces in young and old heads: The factor of age in face recognition. *Psychology and Aging*, *6*, 623–630. doi:10.1037/0882-7974.6.4.623

Funk, F., & Todorov, A. (2013). Criminal stereotypes in the courtroom: Facial tattoos affect guilt and punishment differently. *Psychology, Public Policy, and Law*, *19*, 466–478. doi:10.1037/a0034736

Goodsell, C. A., Gronlund, S. D., & Carlson, C. A. (2010). Exploring the sequential lineup advantage using WITNESS. *Law and Human Behavior*, *34*, 445–459. doi:10.1007/s10979-009-9215-7

Goodsell, C. A., Neuschatz, J. S., & Gronlund, S. D. (2009). Effects of mugshot commitment on lineup performance in young and older adults. *Applied Cognitive Psychology*, *23*, 788–803. doi:10.1002/acp

Goshen-Gottstein, Y., & Groner, L. (2016, July). An ROC analysis of neutral and biased lineup instructions in the discrimination of suspects. In C. Dodson (Chair), *Eyewitness identification: Confidence, accuracy, and justifications.*

Symposium conducted at the International Conference on Memory, Budapest, Hungary.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*, 504–528. doi:10.1016/S0092- 6566(03)00046-1

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* Oxford, England: Wiley.

Gronlund, S. D., Carlson, C. A, Dailey, S. B., & Goodsell, C. A. (2009). Robustness of the sequential lineup advantage. *Journal of Experimental Psychology: Applied*, *15*, 140–152. doi:10.1037/a0015082

Gronlund, S. D., Carlson, C. A., Neuschatz, J. S., Goodsell, C. A., Wetmore, S. A., Wooten, A., & Graham, M. (2012). Showups versus lineups: An evaluation using ROC analysis. *Journal of Applied Research in Memory and Cognition*, *1*, 221–228. doi:10.1016/j.jarmac.2012.09.003

Gronlund, S. D., Mickes, L., Wixted, J. T., & Clark, S. E. (2015). Conducting an eyewitness lineup: How the research got it wrong. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 63, pp. 1–43). New York: Academic Press.

Gronlund, S. D., Wixted, J. T., & Mickes, L. (2014). Evaluating eyewitness identification procedures using receiver operating characteristic analysis. *Current Directions in Psychological Science*, *23*, 3–10. doi:10.1177/0963721413498891

Havard, C., & Memon, A. (2009). The influence of face age on identification from a video line-up: A comparison between older and younger adults. *Memory*, *17*, 847–859. doi:10.1080/09658210903277318

Healy, M. R., Light, L. L., & Chung, C. (2005). Dual-process models of associative recognition in young and older adults: Evidence from receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 768–788. doi:10.1037/0278-7393.31.4.768

Horry, R., Brewer, N., Weber, N., & Palmer, M. A. (2015). The effects of allowing a second sequential lineup lap on choosing and probative value. *Psychology, Public Policy, and Law, 21*, 121–133. doi:10.1037/law0000041

Horry, R., Memon, A., Milne, R., Wright, D. B., & Dalton, G. (2013). Video identification of suspects: A discussion of current practice and policy in the United Kingdom. *Policing*, *7*, 1–9. doi:10.1093/police/pat008

Horry, R., Palmer, M. A., & Brewer, N. (2012). Backloading in the sequential lineup prevents within-lineup criterion shifts that undermine eyewitness identification performance. *Journal of Experimental Psychology: Applied*, *18*, 346–360. doi:10.1037/a0029779

Innocence Project. (2016). *Eyewitness misidentification*. Retrieved from http://www.innocenceproject.org/causes/eyewitness-misidentification/

Jacoby, L. L. (1999). Ironic effects of repetition: Measuring age-related differences in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 3–22. doi:10.1037/0278-7393.25.1.3

Jennings, J. M., & Jacoby, L. L. (1997). An opposition procedure for detecting age-related deficits in recollection: Telling effects of repetition. *Psychology and Aging*, *12*, 352–361. doi:10.1037/0882-7974.12.2.352

Johnston, R. A., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory*, *17*, 577–596. doi:10.1080/09658210902976969

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence–accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1304–1316. doi:10.1037/0278-7393.22.5.1304

Kebbell, M. R., & Milne, R. (1998). Police officers' perceptions of eyewitness performance in forensic investigations. *Journal of Social Psychology*, *138*, 323–330. doi:10.1080/00224549809600384

Key, K. N., Cash, D. K., Neuschatz, J. S., Price, J. L., Wetmore, S. A., & Gronlund, S. D. (2015). Age differences (or lack thereof) in discriminability for lineups and showups. *Psychology, Crime & Law*, *21*, 871–889. doi:10.1080/1068316X.2015.1054387

Koehnken, G., Malpass, R. S., & Wogalter, M. S. (1996). Forensic applications of lineup research. In S. L. Sporer., R. S. Malpass, & G. Koehnken (Eds.), *Psychological issues in eyewitness identification* (pp. 205–231). Mahwah, NJ: Erlbaum.

Koutstaal, W. (2006). Flexible remembering. *Psychonomic Bulletin & Review*, *13*, 84–91. doi:10.3758/BF03193817

Kuehn, L. L. (1974). Looking down a gun barrel: Person perception and violent crime. *Perceptual & Motor Skills, 39*, 1159–1164. Retrieved from http://pms.sagepub.com/

Lacy, J. W., & Stark, C. E. L. (2013). The neuroscience of memory: Implications for the courtroom. *Nature Reviews Neuroscience, 14,* 649–658. doi:10.1038/nrn3563

Lamont, A. C., Stewart-Williams, S., & Podd, J. (2005). Face recognition and aging: Effects of target age and memory load. *Memory & Cognition*, *33*, 1017–1024. doi:10.3758/BF03193209

Lampinen, J. M. (2016). ROC analyses in eyewitness identification research. *Journal of Applied Research in Memory and Cognition, 5,* 21–33. doi:10.1016/j.jarmac.2015.08.006

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174. doi:10.2307/2529310

Lewicki, P., Czyzewska, M., & Hoffman, H. (1987). Unconscious acquisition of complex procedural knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 523–530. doi:10.1037/0278-7393.13.4.523

Lewicki, P., Hill, T., & Czyzewska, M. (1992). Nonconscious acquisition of information. *American Psychologist*, *47*, 796–801. doi:10.1037/0003-066X.47.6.796

Lindsay, D. S., Read, J. D., & Sharma, K. (1998). Accuracy and confidence in person identification: The relationship is strong when witnessing conditions vary widely. *Psychological Science*, *9*, 215–218. doi:10.1111/1467-9280.00041

Lindsay, R. C. L., Martin, R., & Webber, L. (1994). Default values in eyewitness descriptions: A problem for the match-to-description lineup foil selection strategy. *Law and Human Behavior*, *18*, 527–541. doi:10.1007/BF01499172

Lindsay, R. C. L., Wallbridge, H., & Drennan, D. (1987). Do the clothes make the man? An exploration of the effect of lineup attire on eyewitness identification

accuracy. *Canadian Journal of Behavioural Science*, *19*, 463–478. Retrieved from https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=116390

Lindsay, R. C. L., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential presentation. *Journal of Applied Psychology*, *70*, 556–564. doi:10.1037//0021-9010.70.3.556

MacLin, M. K., & Herrera, V. (2006). The criminal stereotype. *North American Journal of Psychology, 8*, 197–208. Retrieved from https://www.questia.com/library/journal/1G1-159922609/the-criminal-stereotype

MacLin, M. K., MacLin, O. H., & Albrechtsen, J. S. (2006). Using image manipulation to construct fair lineups: The case of the Buddy Holly glasses. *Canadian Journal of Police and Security Services*, *4*, 1–16. Retrieved from https://www.questia.com/library/journal/1G1-147390789/using-image-manipulation-to-construct-fair-lineups

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

Malpass, R. S., & Devine, P. G. (1981). Eyewitness identification: Lineup instructions and the absence of the offender. *Journal of Applied Psychology, 66,* 482–489. doi:10.1037/0021-9010.66.4.482

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87,* 252–271. doi:10.1037/0033-295X.87.3.252

Mansour, J. K., Beaudry, J. L., Kalmet, N., Bertrand, M. I., & Lindsay, R. C. L. (in press). Evaluating lineup fairness: Variations across methods and measures. *Law and Human Behavior*. Advance online publication. doi:10.1037/lhb0000203

Marteau, T., & Bekker, H. (1992). The development of a six-item short-form of the state scale of the Spielberger State–Trait Anxiety Inventory (STAI). *British Journal of Psychology*, *31*, 301–306. doi:10.1111/j.2044-8260.1992.tb00997.x

Maurer, D., Grand, R. L., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences, 6,* 255–260. doi:10.1016/S1364-6613(02)01903-4

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*, 3–35. doi:10.1037/1076-8971.7.1.3

Meissner, C. A., Sporer, S. L., & Schooler, J. W. (2007). Person descriptions as eyewitness evidence. In R. C. L. Lindsay., D. E. Ross., J. D. Read, & M. P. Toglia (Eds.), *Handbook of eyewitness psychology: Memory for people* (Vol. 2, pp. 1–34), Mahwah, NJ: Erlbaum.

Memon, A., & Gabbert, F. (2003). Improving the identification accuracy of senior witnesses: Do prelineup questions and sequential testing help? *The Journal of Applied Psychology*, *88*, 341–347. doi:10.1037/0021-9010.88.2.341

Memon, A., Hope, L., Bartlett, J., & Bull, R. (2002). Eyewitness recognition errors: The effects of mugshot viewing and choosing in young and old adults. *Memory & Cognition*, *30*, 1219–1227. doi:10.3758/BF03213404

Metzger, M. M. (1999). Differential effects of transformation on facial recognition in young children: A pilot study. *Perceptual and Motor Skills, 89,* 799–807. doi:10.2466/PMS.89.7.799-807

Metzger, M. M. (2001). Which transformations of stimuli are the most disruptive to facial recognition? *Perceptual and Motor Skills*, *92*, 517–526. doi:10.2466/pms.2001.92.2.517

Mickes, L. (2015). Receiver operating characteristic analysis and confidence-accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition*, *4*, 93–102. doi:10.1016/j.jarmac.2015.01.003

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology: Applied*, *18*, 361–376. doi:10.1037/a0030609

Mickes, L., Hwe, V., Wais, P. E., & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology: General*, *140*, 239–257. doi:10.1037/a0023007

Mickes, L., Moreland, M. B., Clark, S. E., & Wixted, J. T. (2014). Missing the information needed to perform ROC analysis? Then compute *d′*, not the

diagnosticity ratio. *Journal of Applied Research in Memory and Cognition*, *3*, 58–62. doi:10.1016/j.jarmac.2014.04.007

Molinaro, P. F., Arndorfer, A., & Charman, S. D. (2013). Appearance-change instruction effects on eyewitness lineup identification accuracy are not moderated by amount of appearance change. *Law and Human Behavior*, *37*, 432–440. doi:10.1037/lhb0000049

Moreland, M. B. (2015). *Decision processes in eyewitness identification*. Unpublished doctoral dissertation, University of California, Riverside, California.

National Institute of Justice. (2016). *Mission of the National Institute of Justice.* Retrieved from http://www.nij.gov/about/Pages/welcome.aspx

National Research Council (2014). *Identifying the culprit: Assessing eyewitness identification*. Washington, DC: The National Academies Press.

Navon, D. (1992). Selection of lineup foils by similarity to the suspect is likely to misfire. *Law and Human Behavior, 16,* 575–593. doi:10.1007/ BF01044624

Nelson, T. O. (1978) Detecting small amounts of information in memory: Savings for nonrecognized items. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 453–468. doi:10.1037/0278-7393.4.5.453

Neuschatz, J. S., Preston, E. L., Burkett, A. D., Toglia, M. P., Lampinen, J. M., Neuschatz, J. S., . . . Goodsell, C. A. (2005). The effects of post-identification feedback and age on retrospective eyewitness memory. *Applied Cognitive Psychology*, *19*, 435–453. doi:10.1002/acp.1084

Neuschatz, J. S., Wetmore, S. A., Key, K. N., Cash, D. K., Gronlund, S. D., & Goodsell, C. A. (2016). A comprehensive evaluation of showups. In M. K. Miller & B. H. Bornstein (Eds.), *Advances in psychology and law* (Vol. 1, pp. 43–70). doi:10.1007/978-3-319-29406-3_2

Niblett v. Commonwealth, 217 Va. 76, 81, 225 S.E.2d 391, 394 (1976).

Palmer, M. A., & Brewer, N. (2012). Sequential lineup presentation promotes less-biased criterion setting but does not improve discriminability. *Law and Human Behavior*, *36*, 247–255. doi:10.1037/h0093923

Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of

exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied*, *19*, 55–71. doi:10.1037/a0031602

Patterson, K. E., & Baddeley, A. D. (1977). When face recognition fails. *Journal of Experimental Psychology: Human Learning and Memory*, *3*, 406–417. doi:10.1037/0278-7393.3.4.406

Penrod, S. D., & Cutler, B. (1995). Witness confidence and witness accuracy: Assessing their forensic relation. *Psychology, Public Policy, and Law*, *1*, 817–845. doi:10.1037/1076-8971.1.4.817

People v. Adams, 137 Cal. App. 3d 346 (Cal. Ct. App. 1982).

People v. Bethea, 18 Cal. App. 3d 930, 938 (Cal. Ct. App. 1971).

People v. Gonzalez, 38 Cal. 4th 932, 944 (Cal. 2006).

Police and Criminal Evidence Act 1984, Code D. (2011). Retrieved from https://www.gov.uk/police-and-criminal-evidence-act-1984-pace-codes-of-practice

Porter, D., Moss, A., & Reisberg, D. (2014). The appearance-change instruction does not improve line-up identification accuracy. *Applied Cognitive Psychology*, *28*, 151–160. doi:10.1002/acp.2985

Pozzulo, J. D., Lemieux, J. M. T., Wells, E., & McCuaig, H. J. (2006). The influence of eyewitness identification decisions and age of witness on jurors' verdicts and perceptions of reliability. *Psychology, Crime & Law, 12,* 641–652. doi:10.1080/10683160500415539

Pozzulo, J. D., Lemieux, J. M. T., Wilson, A., Crescini, C., & Girardi, A. (2009). The influence of identification decision and DNA evidence. *Journal of Applied Social Psychology*, *39*, 2069–2088. doi:10.1111/j.1559-1816.2009.00516.x

R Development Core Team. (2015). R: A language and environment for statistical computing (Version 3.2.0) [Computer software]. Retrieved from https://www.r-project.org/index.html

Read, J. D. (1995). The availability heuristic in person identification: The sometimes misleading consequences of enhanced contextual information. *Applied Cognitive Psychology*, *9*, 91–121. doi:10.1002/acp.2350090202

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and

compare ROC curves. *BMC Bioinformatics*, *12*, 77–84. doi:10.1186/1471-2105-12-77

Roediger, H. L., Wixted, J. T., & DeSoto, K. A. (2012). The curious complexity between confidence and accuracy in reports from memory. In L. Nadel & W. Sinnott-Armstrong (Eds.), *Memory and law* (pp. 84–117). doi:10.1093/acprof:oso/9780199920754.003.0004

Rose, R., Bull, R., & Vrij, A. (2005). Non-biased lineup instructions do matter - a problem for older witnesses. *Psychology, Crime & Law*, *11*, 147–159. doi:10.1080/10683160512331316307

RStudio Team. (2015). RStudio: Integrated development for R (Version 0.98.1103) [Computer software]. Retrieved from http://www.rstudio.com/

Sampaio, C., & Brewer, W. F. (2009). The role of unconscious memory errors in judgments of confidence for sentence recognition. *Memory & Cognition, 37,* 158–163. doi:10.3758/MC.37.2.158

Sauer, J. D., Brewer, N., & Weber, N. (2008). Multiple confidence estimates as indices of eyewitness memory. *Journal of Experimental Psychology: General, 137,* 528–547. doi:10.1037/a0012712

Sauer, J. D., Brewer, N., Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence-accuracy relationship for eyewitness identification. *Law and Human Behavior*, *34*, 337–347. doi:10.1007/s10979-009-9192-x

Sauerland, M., & Sporer, S. L. (2009). Fast and confident: Postdicting eyewitness identification accuracy in a field study. *Journal of Experimental Psychology: Applied*, *15*, 46–62. doi:10.1037/a0014560

Seale-Carlisle, T. M., & Mickes, L. (2016). US line-ups outperform UK line-ups. *Royal Society Open Science*. doi:10.1098/rsos.160300

Searcy, J. H., Bartlett, J. C., & Memon, A. (1999). Age differences in accuracy and choosing in eyewitness identification and face recognition. *Memory & Cognition*, *27*, 538–552. doi:10.3758/BF03211547

Searcy, J. H., Bartlett, J. C., & Memon, A. (2000). Influence of post-event narratives, line-up conditions and individual differences on false identification by young and older eyewitnesses. *Legal and Criminological Psychology*, *5*, 219–235. doi:10.1348/135532500168100

Searcy, J. H., Bartlett, J. C., Memon, A., & Swanson, K. (2001). Aging and lineup performance at long retention intervals: Effects of metamemory and context reinstatement. *The Journal of Applied Psychology*, *86*, 207–214. doi:10.1037/0021-9010.86.2.207

Shapiro, P. N., & Penrod, S. D. (1986). Meta-analysis of facial identification studies. *Psychological Bulletin*, *100*, 139–156. doi:10.1037/0033-2909.100.2.139

Shaw, J. S., & McClure, K. A. (1996). Repeated postevent questioning can lead to elevated levels of eyewitness confidence. *Law and Human Behavior, 20,* 629–653. doi:10.1007/BF01499235

Sherman, L. W. (1998). *Evidence-based policing.* Washington, DC: Police Foundation. Retrieved from https://www.policefoundation.org/ideas-in-american-policing/

Smith, A. M., Wells, G. L., Lindsay, R. C. L., & Penrod, S. D. (2016). Fair lineups are better than biased lineups and showups, but not because they increase underlying discriminability. *Law and Human Behavior*. Advance online publication. doi:10.1037/lhb0000219

Sporer, S. L., & Martschuk, N. (2014). The reliability of eyewitness identifications by the elderly: An evidence-based review. In M. P. Toglia., D. F. Ross., J. Pozzulo, & E. Pica (Eds.), *The elderly eyewitness in court* (pp. 3–37). New York: Psychology Press.

Sporer, S. L., Penrod, S. D., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin*, *118*, 315–327. doi:10.1037/0033-2909.118.3.315

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. C. L. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law and Human Behavior*, *25*, 459–473. doi:10.1023/A:1012888715007

Steblay, N. K., Dysart, J. E., & Wells, G. L. (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. *Psychology, Public Policy, and Law*, *17*, 99–139. doi:10.1037/a0021650

Steblay, N. M. (1997). Social influence in eyewitness recall: A meta-analytic review of lineup instruction effects. *Law and Human Behavior, 21,* 283–297. doi: 10.1023/A:1024890732059

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1379–1396. doi:10.1037/0278-7393.24.6.1379

Tanaka, J. W., & Gordon, I. (2011). Features, configuration, and holistic face processing. In G. Rhodes., A. Calder., M. Johnson, & J. V. Haxby (Eds.), *Oxford handbook of face perception* (pp. 149–176). doi:10.1093/oxfordhb/9780199559053.013.0010

Technical Working Group for Eyewitness Evidence. (1999). *Eyewitness evidence: A guide for law enforcement*. Washington, DC: U.S. Department of Justice, Office of Justice Programs. Retrieved from https://www.ncjrs.gov/pdffiles1/nij/178240.pdf

Technical Working Group for Eyewitness Evidence. (2003). *Eyewitness evidence: A trainer's manual for law enforcement.* Washington, DC: U.S. Department of Justice, Office of Justice Programs. Retrieved from http://www.nij.gov/publications/Pages/publication-detail.aspx?ncjnumber=188678

Terry, R. L. (1994). Effects of facial transformations on accuracy of recognition. *The Journal of Social Psychology*, *134*, 483–492. doi:10.1080/00224545.1994.9712199

Tredoux, C. G. (1999). Statistical considerations when determining measures of lineup size and lineup bias. *Applied Cognitive Psychology, 13,* S9–S26. doi:10.1002/(SICI)1099-0720(199911)13:1+<S9::AID-ACP634>3.0.CO;2-1

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*, 352–373. doi:10.1037/h0020071

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion and race in face recognition. *Quarterly Journal of Experimental Psychology, 43A,* 161–204. doi:10.1080/14640749108400966

Valentine, T., & Heaton, P. (1999). An evaluation of the fairness of police line-ups and video identifications. *Applied Cognitive Psychology*, *13*, S59–S72. doi:10.1002/(SICI)1099- 0720(199911)13:1+<S59::AID-ACP679>3.0.CO;2-Y

Valentine, T., Hughes, C., & Munro, R. (2009). Recent developments in eyewitness identification procedures in the United Kingdom. In R. Bull., T. Valentine, & T. Williamson (Eds.), *Handbook of psychology of investigative interviewing: Current developments and future directions* (pp. 221–240). doi:10.1002/9780470747599.ch13

van Koppen, P., & Lochun, S. (1997). Portraying perpetrators: The validity of offender descriptions by witnesses. *Law and Human Behavior, 21*, 661–685. doi:10.1023/A:1024812831576

Weber, N., & Brewer, N. (2004). Confidence-accuracy calibration in absolute and relative face recognition judgments. *Journal of Experimental Psychology: Applied, 10,* 156–172. doi:10.1037/ 1076-898X.10.3.156

Wells, G. L. (1978). Applied eyewitness-testimony research: System variables and estimator variables. *Journal of Personality and Social Psychology*, *36*, 1546–1557. doi:10.1037/0022-3514.36.12.1546

Wells, G. L. (1984). The psychology of lineup identifications. *Journal of Applied Social Psychology*, *14*, 89–103. doi:10.1111/j.1559-1816.1984.tb02223.x

Wells, G. L. (1993). What do we know about eyewitness identification? *American Psychologist*, *48*, 553–571. doi: 10.1037/0003-066X.48.5.577

Wells, G. L. (2001). Eyewitness lineups: Data, theory, and policy. *Psychology, Public Policy, and Law, 7,* 791–801. doi:10.1037/1076-8971.7.4.791

Wells, G. L., & Bradfield, A. L. (1999). Distortions in eyewitnesses' recollections: Can the postidentification-feedback effect be moderated*? Psychological Science, 10,* 138–144. doi:10.1111/1467-9280.00121

Wells, G. L., Leippe, M. R., & Ostrom, T. M. (1979). Guidelines for empirically assessing the fairness of a lineup. *Law and Human Behavior*, *3*, 285–293. doi:10.1007/BF01039807

Wells, G. L., & Lindsay, R. C. L. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin*, *88*, 776–784. doi:10.1037/0033-2909.88.3.776

Wells, G. L., Lindsay, R. C. L., & Ferguson, T. J. (1979). Accuracy, confidence, and juror perceptions in eyewitness identification. *Journal of Applied Psychology*, *64*, 440–448. doi:10.1037/0021-9010.64.4.440

Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence: Improving its probative value. *Psychological Science in the Public Interest, 7*, 45–75. doi:10.1111/j.1529-1006.2006.00027.x

Wells, G. L., Rydell, S. M., & Seelau, E. P. (1993). The selection of distractors for eyewitness lineups. *Journal of Applied Psychology*, *78*, 835–844. doi:10.1037/0021-9010.78.5.835

Wells, G. L., Smalarz, L., & Smith, A. M. (2015). ROC analysis of lineups does not measure underlying discriminability and has limited value. *Journal of Applied Research in Memory and Cognition*, *4*, 313–317. doi:10.1016/j.jarmac.2015.08.008

Wells, G. L., Small, M., Penrod, S. D., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior, 22,* 603–647. doi:10.1023/A:1025750605807

Wells, G. L., Steblay, N. K., & Dysart, J. E. (2015). Double-blind photo lineups using actual eyewitnesses: An experimental test of a sequential versus simultaneous lineup procedure. *Law and Human Behavior*, *39*, 1–14. doi:10.1037/lhb0000096

Wetmore, S. A., Neuschatz, J. S., Gronlund, S. D., Wooten, A., Goodsell, C. A., & Carlson, C. A. (2015). Effect of retention interval on showup and lineup performance. *Journal of Applied Research in Memory and Cognition*, *4*, 8–14. doi:10.1016/j.jarmac.2014.07.003

Wilcock, R., Bull, R., & Vrij, A. (2005). Aiding the performance of older eyewitnesses: Enhanced non-biased line-up instructions and line-up presentation. *Psychiatry, Psychology and Law*, *12*, 129–140. doi:10.1375/pplt.2005.12.1.129

Willoughby, M. (2015). *Witnessing crime – Findings from the crime survey for England and Wales 2013/14.* London, England: Ministry of Justice. Retrieved from https://www.gov.uk/government/publications/witnessing-crime-findings-from-the-crime-survey-2013-to-2014

Winograd, E. (1981). Elaboration and distinctiveness in memory for faces. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 181–190. doi:10. 1037/0278-7393.7.3.181

Wixted, J. T., Gronlund, S. D., & Mickes, L. (2014). Policy regarding the sequential lineup is not informed by probative value but is informed by receiver operating characteristic analysis. *Current Directions in Psychological Science*, *23*, 17–18. doi:10.1177/0963721413510934

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review*, *121*, 262–276. doi:10.1037/a0035940

Wixted, J. T., & Mickes, L. (2015). ROC analysis measures objective discriminability for any eyewitness identification procedure. *Journal of Applied Research in Memory and Cognition*, *4*, 329–334. doi:10.1016/j.jarmac.2015.08.007

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger, H. L. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist, 70*, 515–526. doi:10.1037/a0039510

Wixted, J. T., Mickes, L., Dunn, J. C., Clark, S. E., & Wells, W. (2016). Estimating the reliability of eyewitness identifications from police lineups. *Proceedings of the National Academy of Sciences*, *113*, 304–309. doi:10.1073/pnas.1516814112

Wixted, J. T., Read, J. D., & Lindsay, D. S. (2016). The effect of retention interval on the eyewitness identification confidence–accuracy relationship. *Journal of Applied Research in Memory and Cognition, 5,* 192–203. doi:10.1016/j.jarmac.2016.04.006

Wixted, J. T., & Wells, G. L. (2016). *The relationship between eyewitness confidence and identification accuracy: A new synthesis*. Manuscript submitted for publication.

Wogalter, M. S., Malpass, R. S., & Mcquiston, D. E. (2004). A national survey of US police on preparation and conduct of identification lineups. *Psychology, Crime & Law*, *10*, 69–82. doi:10.1080/10683160410001641873

World Health Organization (2015, September). *Ageing and health.* Retrieved from http://www.who.int/topics/ageing/en/

Wright, D. B. (2006). Causal and associative hypotheses in psychology: Examples from eyewitness testimony research. *Psychology, Public Policy, and Law*, *12*, 190–213. doi:10.1037/1076-8971.12.2.190

Wylie, L. E., Bergt, S., Haby, J., Brank, E. M., & Bornstein, B. H. (2015). Age and lineup type differences in the own-race bias. *Psychology, Crime & Law*, *21*, 490–506. doi:10.1080/1068316X.2014.989173

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*, 441–517. doi:10.1006/jmla.2002.2864

Young, A. W., Hay, D. C., McWeeny, K. H., Flude, B. M., & Ellis, A. W. (1985). Matching familiar and unfamiliar faces on internal and external features. *Perception, 14,* 737–746. doi:10.1068/p140737

Zarkadi, T. (2009). *Eyewitness identification: Improving police lineups for suspects with distinctive features* (Doctoral dissertation, University of Warwick, England). Retrieved from http://wrap.warwick.ac.uk/id/eprint/3834

Zarkadi, T., Wade, K. A., & Stewart, N. (2009). Creating fair lineups for suspects with distinctive features. *Psychological Science*, *20*, 1448–1453. doi:10.1111/j.1467-9280.2009.02463.x

Appendices

Appendix A: Modelling in Chapter 2

We fit a signal detection process model of lineup performance to our data (Wixted & Mickes, 2014; see Chapter 1 for a description of the model). Recall that in the do-nothing target-absent lineups, the innocent suspect had a similar distinctive feature to the culprit, but the other foils did not. Therefore, the model for the unfair (do-nothing) lineups consisted of three memory strength distributions (μ_{guilty}, $\mu_{innocent}$, and μ_{foil}). In the fair (replication, pixelation and block) target-absent lineups, there was no designated innocent suspect who was more similar to the culprit than the other foils (i.e., $\mu_{innocent} = \mu_{foil}$), so the model for the fair lineups consisted of two memory strength distributions (μ_{guilty} and $\mu_{innocent}$). For each lineup type, we measured the distance between the μ_{guilty} and $\mu_{innocent}$ distributions (d'), which reflects underlying theoretical discriminability—subjects' ability to discriminate between innocent and guilty suspects. Lower values of d' reflect a poorer ability to discriminate between innocent and guilty suspects and correspond to ROC curves that are closer to the diagonal chance line.

The model assumes that there is a set of response criteria that reflect different levels of confidence. We collapsed our data to a 5-point confidence scale to limit the number of parameters. We combined confidence ratings of 0–20 (c_1), 30–40 (c_2), 50–60 (c_3), 70–80 (c_4), and 90–100 (c_5). The model assumes that an identification is made when the most familiar face in the lineup exceeds c_1—the lowest decision criterion. If no face in the lineup is familiar enough to exceed c_1, the lineup is rejected. The confidence in the identification is determined by the highest criterion that is exceeded.

The unfair lineup model had 7 parameters (μ_{guilty}, μ_{foil}, c_1, c_2, c_3, c_4, c_5), because we fixed $\mu_{innocent}$ to 0 and set the standard deviations for each distribution to 1, for simplicity. For the do-nothing data, target-present lineups had 10 degrees of freedom because there were 5 levels of confidence for guilty suspect identifications and 5 levels of confidence for foil identifications (once those 10 values were known, the number of target-present lineup rejections was fixed). Do-nothing target-absent lineups also had 10 degrees of freedom because there were 5 levels of confidence for innocent suspect identifications and 5 levels of confidence for foil identifications

(with the number of target-absent lineup rejections again fixed). Thus, there were 10 + 10 = 20 degrees of freedom in the data. Because the unfair lineup model had 7 free parameters, the fit of this model to the data involved 20 – 7 = 13 degrees of freedom.

The fair lineup model (which was used for the replication, pixelation and block conditions) had 6 parameters (μ_{guilty}, c_1, c_2, c_3, c_4, c_5). Again, we fixed $\mu_{innocent}$ to 0 and set the standard deviations for each distribution to 1. For fair lineup data, target-present lineups had 10 degrees of freedom because there were 5 levels of confidence for guilty suspect identifications and 5 levels of confidence for foil identifications. For fair target-absent lineups, there were only 5 degrees of freedom because there were 5 confidence criteria for any foil/innocent suspect identifications. Thus, there were 10 + 5 = 15 degrees of freedom in the fair data. Because the fair lineup model had 6 free parameters, the fit of this model to the data involved 15 – 6 = 9 degrees of freedom.

We fit the equal-variance model to our data by minimising the chi-square goodness-of-fit statistic. Three separate pairwise comparisons were performed: replication versus do-nothing, pixelation versus do-nothing, and block versus do-nothing. We first fit the model allowing d' to differ across the two conditions being compared. Table A.1 shows our observed data and the values predicted by the best-fitting signal detection model. It is clear that the model was able to capture the basic trends in our data.[10] The left-hand column (full model) in Table A.2 shows the parameters predicted by the best-fitting signal detection model. It is clear that, compared to the fair (replication, pixelation and block) lineups, doing nothing to prevent the distinctive suspect from standing out impairs d'.

To test whether the observed difference in d' for each pairwise comparison was significant, we fit the same signal detection model, but this time we constrained d' to be equal across the fair and unfair lineups (while allowing the confidence criteria to differ across conditions). Table A.2 shows the fits for the full model, in which d' was allowed to differ across conditions, and for the constrained model, in which d' was held constant across conditions. In comparison to the full model, the constrained model provided a significantly worse fit of the data for the replication and do-

[10] When we fit the data to the fair (replication, pixelation, block) and unfair (do-nothing) lineup data separately, the model fit the fair lineup data well (replication $p = .37$; pixelation $p = .37$; block $p = .20$), but significantly deviated from the observed data in the do-nothing condition ($p < .001$).

nothing, χ^2 (1) = 24.32, $p < .001$, pixelation and do-nothing, χ^2 (1) = 15.36, $p < .001$, and block and do-nothing, χ^2 (1) = 20.36, $p < .001$, comparisons. These results indicate that d' is significantly lower in the do-nothing condition compared to each of the fair lineup conditions. In short, the signal detection model fitting exercise and our atheoretical *p*AUC analysis led to the same results: doing nothing to prevent a distinctive suspect from standing out markedly impairs subjects' ability to discriminate between innocent and guilty suspects.

It should be noted that our do-nothing data significantly deviated from the predictions of the equal-variance signal detection model. Given our very large *N*, the relatively poor fit may simply reflect high power to detect even slight deviations from the simple model. It is clear from Table A.1 that the equal-variance model captured the overall trends in our data well. Nevertheless, we also fit an unequal-variance version of the model to our data. The model fit in the do-nothing condition—but not the fair lineup conditions—was significantly improved by allowing for unequal variance. We therefore conducted the same model fitting process described above, but measuring d_a (a standard d'-like discriminability measure that applies to the unequal-variance situation) instead of d' (which assumes equal variances). The basic story remained the same: the model fits were significantly worse when we constrained d_a to be the same across the fair and unfair lineup conditions. Again, these results reinforce the conclusions from our *p*AUC analysis.

Fitting a signal detection process model to our data tells us at least three valuable pieces of information. First, the model-based analyses indicate that the results of our atheoretical *p*AUC analysis map onto measures of underlying theoretical discriminability (cf. Lampinen, 2015). Second, it confirms that our findings are not simply the product of limiting the *p*AUC analysis to a small subset of the do-nothing curve. The modelling and the *p*AUC results are consistent, even though the modelling uses the largest FAR range that a target-absent lineup can support. Third, it rules out the possibility that filler siphoning can account for our results (cf. Wells, Smalarz, & Smith, 2015). Wells, Smalarz, and Smith suggested that fair lineups elicit fewer innocent suspect identifications than unfair lineups because the presence of plausible alternatives (i.e. fair foils) siphon some of the incorrect identifications away from the innocent suspect. Filler siphoning is certainly

present in our data—there is a large difference in willingness to choose the suspect in the fair and unfair lineup conditions (see Figure 2.2 in Chapter 2). However, the signal detection model accounts for filler identifications and yet the model still required that d' differ across the conditions to significantly improve the fit (see also Wixted & Mickes, 2015). This result indicates that filler siphoning cannot explain either the d' or the *p*AUC advantage evident in our fair lineup conditions.

Table A.1
Observed and Predicted Identification Responses in Each Confidence Bin in the Replication, Pixelation, Block, and Do-nothing Lineups

	Target present			Target absent		
Confidence	Guilty suspect	Foil	Incorrect rejection	Innocent suspect	Foil	Correct rejection
			Replication			
0–20						
Observed	21.00	45.00	-	-	57.00	-
Predicted	20.99	38.17	-	-	64.20	-
30–40						
Observed	36.00	52.00	-	-	99.00	-
Predicted	33.96	58.33	-	-	94.28	-
50–60						
Observed	96.00	127.00	-	-	207.00	-
Predicted	89.77	135.01	-	-	203.00	-
70–80						
Observed	106.00	93.00	-	-	168.00	-
Predicted	96.58	113.68	-	-	156.10	-
90–100						
Observed	88.00	65.00	-	-	96.00	-
Predicted	94.42	68.03	-	-	86.49	-
Total						
Observed	-	-	396.00	-	-	513.00
Predicted	-	-	376.05	-	-	535.93
			Pixelation			
0–20						
Observed	29.00	52.00	-	-	89.00	-
Predicted	28.57	54.87	-	-	85.84	-
30–40						
Observed	38.00	56.00	-	-	119.00	-
Predicted	39.16	70.31	-	-	105.31	-
50–60						
Observed	92.00	151.00	-	-	197.00	-
Predicted	92.60	144.61	-	-	202.28	-
70–80						
Observed	84.00	105.00	-	-	131.00	-
Predicted	83.80	102.68	-	-	132.80	-

90–100						
Observed	77.00	47.00	-	-	78.00	-
Predicted	74.24	57.84	-	-	70.43	-
Total						
Observed	-	-	414.00	-	-	510.00
Predicted	-	-	396.31	-	-	527.34
			Block			
0–20						
Observed	27.00	62.00	-	-	89.00	-
Predicted	30.70	56.03	-	-	91.12	-
30–40						
Observed	51.00	61.00	-	-	105.00	-
Predicted	40.82	69.22	-	-	107.38	-
50–60						
Observed	101.00	137.00	-	-	222.00	-
Predicted	100.84	146.07	-	-	210.32	-
70–80						
Observed	71.00	93.00	-	-	133.00	-
Predicted	84.05	92.35	-	-	122.23	-
90–100						
Observed	73.00	37.00	-	-	54.00	-
Predicted	65.25	44.11	-	-	55.04	-
Total						
Observed	-	-	414.00	-	-	534.00
Predicted	-	-	397.55	-	-	550.91
			Do-nothing			
0–20						
Observed	17.00	32.00	-	18.00	29.00	-
Predicted	28.65	19.74	-	26.35	28.04	-
30–40						
Observed	35.00	36.00	-	37.00	50.00	-
Predicted	49.22	30.15	-	42.38	41.12	-
50–60						
Observed	156.00	70.00	-	113.00	69.00	-
Predicted	148.19	68.25	-	110.26	85.91	-
70–80						
Observed	155.00	44.00	-	74.00	49.00	-
Predicted	145.58	41.84	-	87.71	47.48	-
90–100						
Observed	266.00	24.00	-	122.00	22.00	-
Predicted	264.29	28.69	-	106.32	29.32	-
Total						
Observed	-	-	275.00	-	-	434.00
Predicted	-	-	285.40	-	-	412.11

Note. The total row displays all reject identification decisions because the model does not account for the confidence level with which lineup rejections are made.

Table A.2
Full and Constrained (d') Model Fits for the Fair (Replication, Pixelation, Block) vs. Unfair (Do-nothing) Lineup Comparisons

	Full model		Constrained model	
Estimate	Replication	Do-nothing	Replication	Do-nothing
μ_{guilty} (d')	0.86	0.54	0.73	0.73
μ_{foil}	-	–1.25	-	–1.15
c_1	1.18	0.22	1.16	0.32
c_2	1.27	0.31	1.25	0.41
c_3	1.41	0.46	1.39	0.56
c_4	1.76	0.85	1.74	0.95
c_5	2.22	1.25	2.20	1.35
Overall χ^2	49.41		73.73	
Overall df	22		23	
Overall *p*	<.001		<.001	
	Pixelation	Do-nothing	Pixelation	Do-nothing
μ_{guilty} (d')	0.80	0.54	0.69	0.69
μ_{foil}	-	–1.25	-	–1.17
c_1	1.18	0.22	1.17	0.30
c_2	1.30	0.31	1.29	0.39
c_3	1.46	0.46	1.45	0.54
c_4	1.84	0.85	1.82	0.93
c_5	2.30	1.25	2.28	1.33
Overall χ^2	45.91		61.27	
Overall df	22		23	
Overall *p*	.002		<.001	
	Block	Do-nothing	Block	Do-nothing
μ_{guilty} (d')	0.83	0.54	0.72	0.72
μ_{foil}	-	–1.25	-	–1.16
c_1	1.21	0.22	1.19	0.31
c_2	1.34	0.31	1.32	0.40
c_3	1.50	0.46	1.48	0.55
c_4	1.91	0.85	1.89	0.95
c_5	2.40	1.25	2.37	1.34
Overall χ^2	48.28		68.64	
Overall df	22		23	
Overall *p*	.001		<.001	

Note. The full model allows d' to differ between the two lineups being compared. The constrained model holds d' constant across the two lineups being compared. Overall χ^2, df and *p* rows represent goodness-of-fit statistics when the model was fit to the two lineups together.

Appendix B: Confidence and accuracy in Chapter 2

Table B.1
Frequencies of Identification Responses in Each Confidence Bin in the Replication, Pixelation, Block, and Do-nothing Lineups

Lineup type and confidence	Target present			Target absent		
	Guilty suspect	Foil	Incorrect rejection	Innocent suspect	Foil	Correct rejection
Replication						
0–20	21.00	45.00	40.00	9.50	47.50	40.00
30–40	36.00	52.00	45.00	16.50	82.50	68.00
50–60	96.00	127.00	124.00	34.50	172.50	158.00
70–80	106.00	93.00	102.00	28.00	140.00	120.00
90–100	88.00	65.00	85.00	16.00	80.00	127.00
Pixelation						
0–20	29.00	52.00	30.00	14.83	74.17	41.00
30–40	38.00	56.00	53.00	19.83	99.17	56.00
50–60	92.00	151.00	130.00	32.83	164.17	163.00
70–80	84.00	105.00	106.00	21.83	109.17	134.00
90–100	77.00	47.00	95.00	13.00	65.00	116.00
Block						
0–20	27.00	62.00	41.00	14.83	74.17	59.00
30–40	51.00	61.00	63.00	17.50	87.50	62.00
50–60	101.00	137.00	132.00	37.00	185.00	158.00
70–80	71.00	93.00	97.00	22.17	110.83	143.00
90–100	73.00	37.00	81.00	9.00	45.00	112.00
Do-nothing						
0–20	17.00	32.00	22.00	18.00	29.00	43.00
30–40	35.00	36.00	36.00	37.00	50.00	50.00
50–60	156.00	70.00	88.00	113.00	69.00	122.00
70–80	155.00	44.00	66.00	74.00	49.00	107.00
90–100	266.00	24.00	63.00	122.00	22.00	112.00

Appendix C: Preliminary analyses in Chapter 3

Before collapsing across the three fair lineup techniques (replication, pixelation and block) in our dataset, we conducted preliminary analyses to check that, within each age group, subjects performed similarly on the three fair lineup types. To this end, we examined subjects' identification responses, conducted ROC analysis and fit a signal detection process model to our data (Wixted & Mickes, 2014).

Identification responses

Figure C.1 shows the identification responses made by the young, middle-aged and older adults in (a) target-present and (b) target-absent lineups, as a function of lineup type. The chi-square tests presented in Chapter 3 indicated that, within each age group, performance was the same on the three fair lineups.

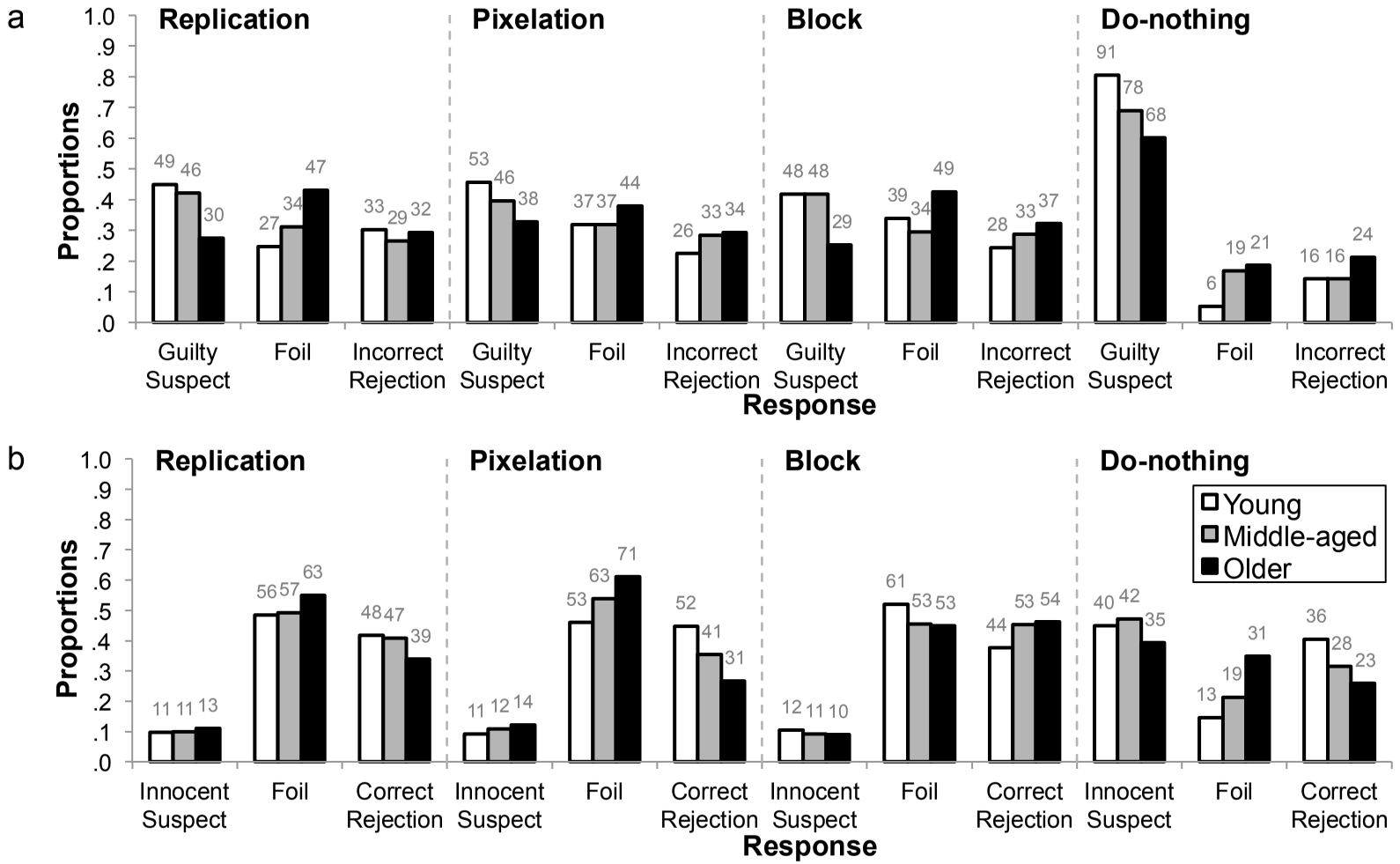

Figure C.1. Identification responses made by the young, middle-aged, and older adults in replication, pixelation, block, and do-nothing (a) target-present and (b) target-absent lineups. Data labels are absolute frequencies.

ROC analysis

To confirm that ability to discriminate between innocent and guilty suspects was the same in the three fair lineups, we conducted ROC analysis. We constructed our ROC curves and calculated our *p*AUC statistics (see Table C.1) in the same way as in Chapter 2. We set the specificity (1 – FAR) to .910, using the FAR range covered by the least extensive curve. Figure C.2 shows the ROC curves for the replication, pixelation, block and do-nothing lineups for (a) young, (b) middle-aged, and (c) older subjects. Within each age group, the ROCs for the replication, pixelation and block lineups lie on top of each other. This indicates that the three fair lineups led to similar levels of identification performance. In young adults, the *p*AUCs did not differ significantly between replication and pixelation ($D = 0.82$, $p = .41$), replication and block ($D = 0.69$, $p = .49$), or block and pixelation ($D = 0.18$, $p = .86$) lineups. Nor did the *p*AUCs differ significantly between replication and pixelation ($D = 0.42$, $p = .67$), replication and block ($D = 0.17$, $p = .86$), or block and pixelation ($D = 0.25$, $p = .80$) lineups in middle-aged adults. Finally, in the older adults, the *p*AUCs did not differ significantly between replication and pixelation ($D = 0.01$, $p = .99$), replication and block ($D = 0.46$, $p = .65$), or block and pixelation ($D = 0.44$, $p = .66$) lineups. Concordant with the analysis of identification responses, this suggests that, within each age group, all three fair techniques were equally effective at enhancing subjects' ability to discriminate between innocent and guilty suspects.

Table C.1
Partial Area Under the Curve (pAUC) Statistics [and 95% Confidence Intervals]

Lineup type	Young	Middle-aged	Older
Replication	0.021 [0.010, 0.035]	0.018 [0.009, 0.032]	0.012 [0.005, 0.021]
Pixelation	0.028 [0.018, 0.040]	0.015 [0.007, 0.027]	0.012 [0.005, 0.022]
Block	0.027 [0.017, 0.037]	0.017 [0.008, 0.029]	0.014 [0.008, 0.023]

Note. Specificity (1 – FAR) = .910, which was set using the FAR range of the least extensive curve.

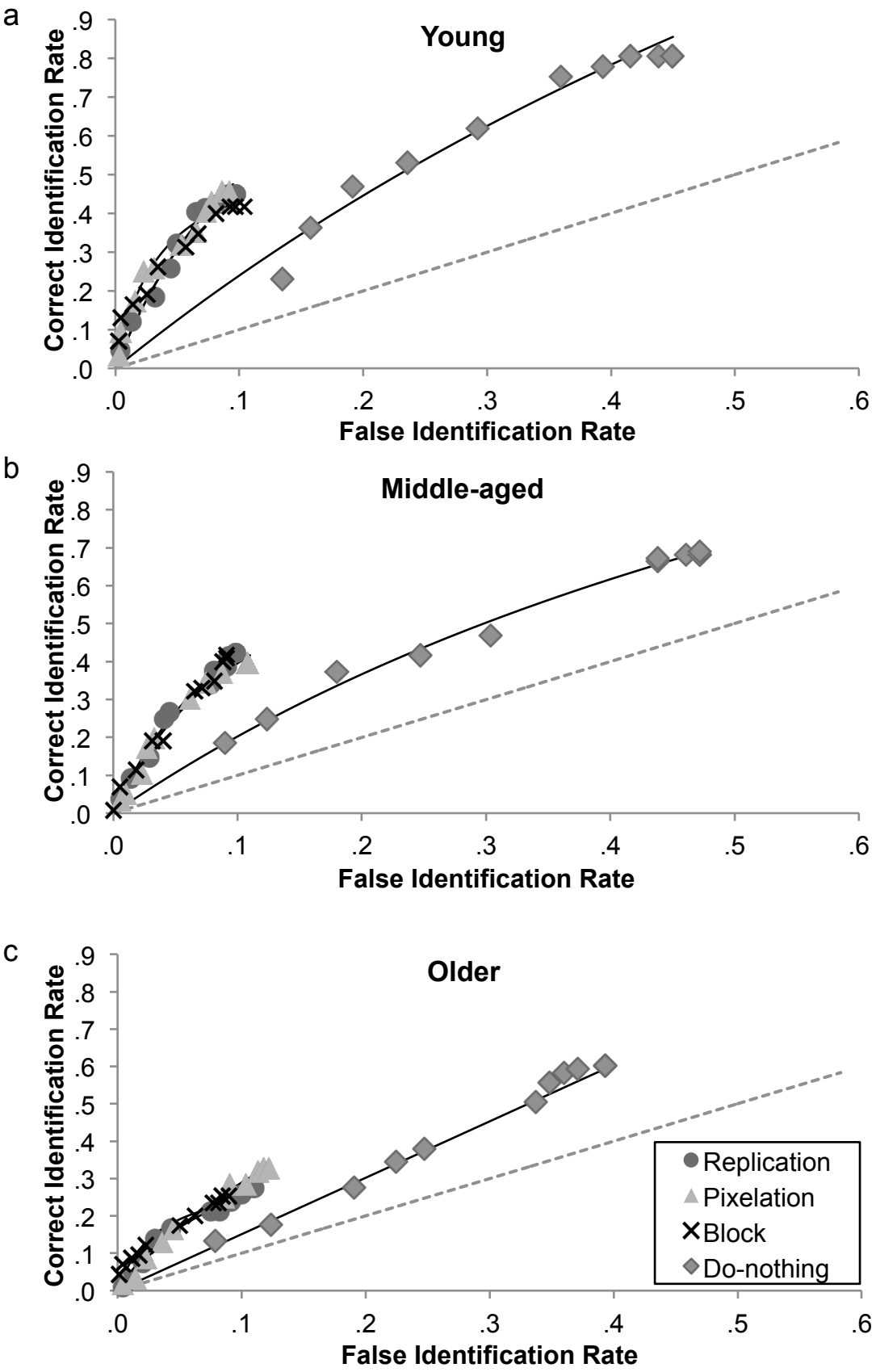

Figure C.2. Receiver operating characteristic (ROC) curves for the replication, pixelation, block, and do-nothing lineups for (a) young, (b) middle-aged, and (c) older adults. The dashed lines represent chance-level performance.

Modelling

To further assess whether all three fair lineup techniques were equally effective, we fit a signal detection model to our data (Wixted & Mickes, 2014; see Chapter 3 for a description of the model). We fit the model to the replication, pixelation and block data in each age group by minimising the chi-square goodness-of-fit statistic. Within each age group, we performed three separate pairwise comparisons: replication versus pixelation, replication versus block, and pixelation versus block. We first fit the model allowing d' to differ across the two conditions being compared (unconstrained model). Table C.2 shows our observed data and the values predicted by the best-fitting signal detection model, while Table C.3 shows the best-fitting parameters and the chi-square goodness-of-fit statistics. It is clear from Table C.2 that the model proficiently captured the trends in our data, and this is reflected in the non-significant (full model) chi-square goodness-of-fit statistics in Table C.3.

To test whether there were any statistically significant differences in d' for each pairwise comparison, we fit the same model, allowing the confidence criteria to differ, but constraining d' to be equal in the two conditions being compared. The overall χ^2, df and p rows in Table C.3 show the full (unconstrained) and constrained model fit statistics. In comparison to the full model, the constrained model did not provide a significantly worse fit of the data for the replication and pixelation (young, χ^2 (1) = 0.18, p = .68; middle-aged, χ^2 (1) = 0.15, p = .70; older, χ^2 (1) = 0.02, p = .89), replication and block (young, χ^2 (1) = 0.01, p = .91; middle-aged, χ^2 (1) = 0.03, p = .86; older, χ^2 (1) = 0.14, p = .71), and pixelation and block (young, χ^2 (1) = 0.31, p = .58; middle-aged, χ^2 (1) = 0.35, p = .55; older, χ^2 (1) = 0.07, p = .79) comparisons. These results indicate that, within each age group, there was no statistically significant difference in d' between the three fair lineup conditions. Overall, our analyses suggest that performance was the same on the three fair lineup types. Therefore, for ease of interpretation, we collapsed the data over the replication, pixelation and block lineups within each age group.

Table C.2

Observed and Predicted Identification Responses in Each Confidence Bin in the Replication, Pixelation, and Block Lineups for the Young, Middle-aged, and Older Adults

	Replication					Pixelation					Block				
	Target present			Target absent		Target present			Target absent		Target present			Target absent	
Confidence	Guilty suspect	Foil	Incorrect rejection	Foil	Correct rejection	Guilty suspect	Foil	Incorrect rejection	Foil	Correct rejection	Guilty suspect	Foil	Incorrect rejection	Foil	Correct rejection
								Young							
0–20															
Observed	3.00	4.00	-	11.00	-	6.00	8.00	-	13.00	-	2.00	7.00	-	16.00	-
Predicted	3.27	4.49	-	9.73	-	5.30	6.83	-	14.80	-	4.54	6.61	-	14.09	-
30–40															
Observed	2.00	7.00	-	11.00	-	10.00	7.00	-	14.00	-	10.00	14.00	-	17.00	-
Predicted	4.26	5.32	-	10.69	-	7.26	8.04	-	15.44	-	8.90	11.06	-	20.77	-
50–60															
Observed	16.00	11.00	-	14.00	-	8.00	6.00	-	21.00	-	14.00	8.00	-	22.00	-
Predicted	10.48	10.95	-	19.63	-	10.52	9.43	-	15.93	-	12.25	11.85	-	19.26	-
70–80															
Observed	15.00	4.00	-	22.00	-	18.00	11.00	-	13.00	-	7.00	6.00	-	15.00	-
Predicted	14.26	10.49	-	16.22	-	16.69	10.14	-	14.79	-	11.00	7.56	-	10.84	-
90–100															
Observed	13.00	1.00	-	9.00	-	11.00	5.00	-	3.00	-	15.00	4.00	-	3.00	-
Predicted	12.92	4.65	-	6.33	-	12.06	3.30	-	4.21	-	12.17	4.39	-	5.70	-
Total															
Observed	-	-	33.00	-	48.00	-	-	26.00	-	52.00	-	-	28.00	-	44.00
Predicted	-	-	27.92	-	52.39	-	-	26.44	-	50.83	-	-	24.68	-	46.33
								Middle-aged							
0–20															
Observed	4.00	6.00	-	5.00	-	3.00	8.00	-	14.00	-	8.00	7.00	-	7.00	-
Predicted	2.83	4.32	-	9.01	-	4.33	7.06	-	13.34	-	4.61	6.52	-	12.82	-
30–40															
Observed	5.00	2.00	-	9.00	-	8.00	7.00	-	18.00	-	3.00	4.00	-	11.00	-
Predicted	3.20	4.55	-	8.91	-	6.66	9.63	-	16.59	-	3.83	4.91	-	8.96	-

50–60															
Observed	10.00	12.00	-	26.00	-	15.00	14.00	-	24.00	-	15.00	9.00	-	24.00	-
Predicted	10.80	13.03	-	22.95	-	13.21	15.29	-	23.32	-	12.38	13.18	-	21.67	-
70–80															
Observed	17.00	8.00	-	18.00	-	14.00	3.00	-	12.00	-	14.00	10.00	-	18.00	-
Predicted	13.25	11.52	-	17.59	-	10.18	8.35	-	11.33	-	14.96	10.72	-	15.27	-
90–100															
Observed	10.00	6.00	-	10.00	-	6.00	5.00	-	7.00	-	8.00	4.00	-	4.00	-
Predicted	12.59	5.72	-	7.75	-	9.08	4.15	-	5.21	-	9.13	3.15	-	4.03	-
Total															
Observed	-	-	29.00	-	47.00	-	-	33.00	-	41.00	-	-	33.00	-	53.00
Predicted	-	-	27.20	-	48.79	-	-	28.07	-	46.22	-	-	31.61	-	54.25
								Older							
0–20															
Observed	4.00	7.00	-	13.00	-	5.00	6.00	-	13.00	-	2.00	10.00	-	9.00	-
Predicted	3.40	7.63	-	12.73	-	3.45	7.72	-	12.59	-	3.46	6.92	-	11.19	-
30–40															
Observed	3.00	6.00	-	11.00	-	4.00	12.00	-	12.00	-	7.00	7.00	-	19.00	-
Predicted	3.06	6.44	-	10.26	-	4.40	9.13	-	14.07	-	5.87	10.73	-	16.37	-
50–60															
Observed	8.00	23.00	-	31.00	-	14.00	13.00	-	35.00	-	9.00	25.00	-	23.00	-
Predicted	11.33	20.53	-	30.32	-	11.71	20.76	-	29.40	-	12.54	18.68	-	26.11	-
70–80															
Observed	12.00	9.00	-	16.00	-	11.00	10.00	-	15.00	-	3.00	6.00	-	9.00	-
Predicted	9.01	12.06	-	16.30	-	8.43	11.48	-	14.91	-	5.23	5.80	-	7.51	-
90–100															
Observed	3.00	2.00	-	5.00	-	4.00	3.00	-	10.00	-	8.00	1.00	-	3.00	-
Predicted	3.51	2.90	-	3.73	-	6.10	5.41	-	6.66	-	5.02	3.71	-	4.63	-
Total															
Observed	-	-	32.00	-	39.00	-	-	34.00	-	31.00	-	-	37.00	-	54.00
Predicted	-	-	29.13	-	41.66	-	-	27.40	-	38.38	-	-	37.05	-	51.19

Note. The total row displays all reject identification decisions because the model does not account for the confidence level with which lineup rejections are made.

Table C.3

Full and Constrained (d') Model Fits for the Replication vs. Pixelation, Replication vs. Block, and Pixelation vs. Block Comparisons in the Young, Middle-aged, and Older Adults

	Young				Middle-aged				Older			
	Full model		Constrained model		Full model		Constrained model		Full model		Constrained model	
Estimate	Replication	Pixelation	Replication	Pixelation	Replication	Pixelation	Replication	Pixelation	Replication	Pixelation	Replication	Pixelation
μ_{guilty} (d')	1.18	1.25	1.20	1.20	1.09	1.02	1.05	1.05	0.70	0.72	0.71	0.71
c_1	1.16	1.13	1.17	1.13	1.11	1.07	1.11	1.08	1.01	0.96	1.01	0.96
c_2	1.30	1.34	1.30	1.33	1.24	1.25	1.23	1.26	1.19	1.14	1.19	1.13
c_3	1.45	1.57	1.46	1.57	1.36	1.49	1.36	1.50	1.33	1.33	1.33	1.33
c_4	1.80	1.89	1.81	1.88	1.74	1.95	1.74	1.96	1.86	1.83	1.86	1.83
c_5	2.35	2.50	2.42	2.55	2.27	2.43	2.27	2.43	2.54	2.33	2.55	2.33
Overall χ^2	24.93		25.11		19.69		19.83		18.75		18.77	
Overall df	18		19		18		19		18		19	
Overall p	.13		.16		.35		.40		.41		.47	
	Replication	Block	Replication	Block	Replication	Block	Replication	Block	Replication	Block	Replication	Block
μ_{guilty} (d')	1.18	1.16	1.17	1.17	1.09	1.12	1.10	1.10	0.70	0.77	0.74	0.74
c_1	1.16	1.07	1.16	1.07	1.11	1.17	1.11	1.17	1.01	1.13	1.02	1.13
c_2	1.30	1.26	1.29	1.26	1.24	1.35	1.24	1.35	1.19	1.28	1.19	1.28
c_3	1.45	1.56	1.45	1.57	1.36	1.48	1.36	1.48	1.33	1.52	1.34	1.52
c_4	1.80	1.96	1.80	1.96	1.74	1.89	1.75	1.88	1.86	2.09	1.87	2.09
c_5	2.35	2.40	2.35	2.40	2.27	2.52	2.27	2.52	2.54	2.47	2.55	2.46
Overall χ^2	29.01		29.02		18.61		18.64		18.14		18.28	
Overall df	18		19		18		19		18		19	
Overall p	.05		.07		.42		.48		.45		.50	
	Pixelation	Block	Pixelation	Block	Pixelation	Block	Pixelation	Block	Pixelation	Block	Pixelation	Block
μ_{guilty} (d')	1.25	1.16	1.20	1.20	1.02	1.12	1.07	1.07	0.72	0.77	0.75	0.75
c_1	1.13	1.07	1.13	1.07	1.07	1.17	1.08	1.17	0.96	1.13	0.96	1.13
c_2	1.34	1.26	1.33	1.27	1.25	1.35	1.26	1.34	1.14	1.28	1.14	1.28
c_3	1.57	1.56	1.56	1.57	1.49	1.48	1.50	1.47	1.33	1.52	1.33	1.52
c_4	1.89	1.96	1.88	1.97	1.95	1.89	1.97	1.88	1.83	2.09	1.83	2.09
c_5	2.50	2.40	2.49	2.41	2.43	2.52	2.43	2.51	2.33	2.47	2.34	2.46
Overall χ^2	18.14		18.45		19.41		19.76		27.78		27.85	
Overall df	18		19		18		19		18		19	
Overall p	.45		.49		.37		.41		.07		.09	

Note. The full (unconstrained) model allows d' to differ between the two lineups being compared. The constrained model holds d' constant across the two lineups being compared. Overall χ^2, df and p rows represent goodness-of-fit statistics when the model was fit to the two lineups together.

Appendix D: Modelling in Chapter 3

To confirm our findings from the ROC analysis, we fit a signal detection model to our data (Wixted & Mickes, 2014). We fit the model to the fair lineups in each age group, and we discuss these findings in detail in Chapter 3. Here, we present the model fits for the remaining comparisons examined in our ROC analysis. That is, we compare performance on the unfair lineups across age groups, and we compare performance on the fair and unfair lineups, within each age group.

The model fit the fair lineup is described in Chapter 3, but the model for an unfair (do-nothing) lineup differs slightly. The model for an unfair lineup consists of three distributions with means of μ_{guilty}, $\mu_{innocent}$, and μ_{foil}. The measure of interest is the distance between the μ_{guilty} and $\mu_{innocent}$ distributions (d'), which, similar to the ROC analysis, reflects the ability to discriminate between guilty and innocent suspects. The unfair lineup data contained 20 degrees of freedom, corresponding to the 5 levels of confidence for guilty suspect identifications and foil identifications in target-present lineups, and the 5 levels of confidence for innocent suspect identifications and foil identifications in target-absent lineups. Once these response frequencies were known, the number of rejections made in target-present and target-absent lineups was fixed. The model had 7 free parameters (μ_{guilty}, μ_{foil}, c_1, c_2, c_3, c_4, c_5) because we fixed $\mu_{innocent}$ to 0 and set the standard deviations for each distribution to 1, for simplicity. Thus, the fit had 20 – 7 = 13 degrees of freedom.[11]

First, we examined how performance changed with age on the unfair lineups. We fit the model to the unfair lineup data in each age group by minimising the chi-square goodness-of-fit statistic. Table D.1 shows our observed unfair data and the values predicted by the best-fitting signal detection model, whereas Table D.2 shows the best-fitting parameters and the chi-square goodness-of-fit statistics. Again, it is clear from Table D.1 that this simple model proficiently captured the trends in our data, and this is reflected in the (full model) chi-square goodness-of-fit statistics in Table D.2.

We performed three separate pairwise comparisons: young versus middle-aged, young versus older, and middle-aged versus older. We fit the same model,

[11] It is more difficult to assume an asymptotic chi-square distribution for cells with fewer than 5 observations (Cochran, 1952). Therefore, we also performed the model fitting, collapsing the data to a 3-point confidence scale. Our results were the same regardless of whether we fit a model with 5 confidence criteria or 3 confidence criteria.

allowing the confidence criteria to differ, but constraining d' to be equal in the two age groups being compared. The overall χ^2, df and p rows in Table D.2 show the full (unconstrained) and constrained model fit statistics. In comparison to the full model, the constrained model did not provide a significantly worse fit of the data for the young and middle-aged, χ^2 (2) = 5.97, p = .05, or the middle-aged and older, χ^2 (2) = 2.88, p = .24, comparisons. However, the constrained model did provide a significantly worse fit of the data for the young and older comparison, χ^2 (2) = 16.20, p < .001. These results support the ROC analysis and suggest that there was no statistically significant difference in d' between the young and middle-aged groups, or the middle-aged and older groups on the unfair lineup. However, the model fitting indicated that d' was significantly worse in the older adults than in the young adults on the unfair lineup (p < .001), but this did not reach statistical significance in the ROC analysis (p = .96). Nevertheless, it is important to note that, regardless of which type of analysis we use, our conclusion remains the same: unfair lineups yield poor discriminability in subjects of all ages.

Table D.1

Observed and Predicted Identification Responses in Each Confidence Bin in the Unfair Lineups for the Young, Middle-aged, and Older Adults

	Target present			Target absent		
Confidence	Guilty suspect	Foil	Incorrect rejection	Innocent suspect	Foil	Correct rejection
			Young			
0–20						
Observed	0.00	2.00	-	3.00	4.00	-
Predicted	3.79	1.35	-	4.05	2.53	-
30–40						
Observed	6.00	1.00	-	5.00	1.00	-
Predicted	5.08	1.48	-	4.78	2.56	-
50–60						
Observed	25.00	1.00	-	11.00	4.00	-
Predicted	19.13	3.54	-	13.79	5.20	-
70–80						
Observed	19.00	2.00	-	7.00	2.00	-
Predicted	16.35	1.50	-	8.31	1.81	-
90–100						
Observed	41.00	0.00	-	14.00	2.00	-
Predicted	42.62	1.02	-	11.32	1.02	-
Total						
Observed	-	-	16.00	-	-	36.00
Predicted	-	-	17.15	-	-	33.62
			Middle-aged			
0–20						
Observed	1.00	3.00	-	1.00	4.00	-
Predicted	2.83	1.90	-	2.74	2.76	-
30–40						
Observed	2.00	2.00	-	2.00	7.00	-
Predicted	4.80	2.79	-	4.27	3.82	-
50–60						
Observed	28.00	9.00	-	17.00	4.00	-
Predicted	23.31	8.77	-	16.47	10.49	-
70–80						
Observed	19.00	5.00	-	11.00	2.00	-
Predicted	18.89	3.47	-	9.77	3.50	-
90–100						
Observed	28.00	0.00	-	11.00	2.00	-
Predicted	28.05	1.60	-	9.14	1.42	-
Total						
Observed	-	-	16.00	-	-	28.00
Predicted	-	-	16.59	-	-	24.64
			Older			
0–20						
Observed	2.00	4.00	-	3.00	11.00	-
Predicted	6.31	5.25	-	5.28	6.10	-
30–40						
Observed	9.00	3.00	-	2.00	6.00	-
Predicted	6.61	4.53	-	5.04	4.95	-
50–60						
Observed	18.00	10.00	-	10.00	11.00	-
Predicted	18.45	8.94	-	12.14	9.02	-
70–80						
Observed	19.00	3.00	-	9.00	3.00	-
Predicted	16.34	4.49	-	8.76	4.11	-
90–100						
Observed	20.00	1.00	-	11.00	0.00	-
Predicted	19.93	2.03	-	7.70	1.71	-
Total						
Observed	-	-	24.00	-	-	23.00
Predicted	-	-	20.13	-	-	24.20

Note. The total row displays all reject identification decisions because the model does not account for the confidence level with which lineup rejections are made.

Table D.2
Full and Constrained (d') Model Fits for the Young vs. Middle-Aged, Young vs. Older, and Middle-aged vs. Older Unfair Lineup Comparisons

	Full model		Constrained model	
Estimate	Young	Middle-aged	Young	Middle-aged
μ_{guilty} (d')	0.83	0.59	0.70	0.70
μ_{foil}	–1.67	–1.44	–1.53	–1.53
c_1	–0.03	–0.12	0.01	–0.15
c_2	0.12	0.00	0.14	–0.02
c_3	0.28	0.17	0.29	0.15
c_4	0.76	0.78	0.77	0.78
c_5	1.14	1.26	1.13	1.27
Overall χ^2	33.05		39.02	
Overall df	26		28	
Overall p	.16		.08	
	Young	Older	Young	Older
μ_{guilty} (d')	0.83	0.43	0.62	0.62
μ_{foil}	–1.67	–1.29	–1.43	–1.43
c_1	–0.03	–0.05	0.03	–0.09
c_2	0.12	0.19	0.16	0.16
c_3	0.28	0.39	0.31	0.36
c_4	0.76	0.88	0.77	0.88
c_5	1.14	1.36	1.13	1.38
Overall χ^2	32.95		49.15	
Overall df	26		28	
Overall p	.16		.008	
	Middle-aged	Older	Middle-aged	Older
μ_{guilty} (d')	0.59	0.43	0.51	0.51
μ_{foil}	–1.44	–1.29	–1.36	–1.36
c_1	–0.12	–0.05	–0.09	–0.07
c_2	0.00	0.19	0.03	0.17
c_3	0.17	0.39	0.19	0.37
c_4	0.78	0.88	0.79	0.87
c_5	1.26	1.36	1.26	1.36
Overall χ^2	36.60		39.48	
Overall df	26		28	
Overall p	.08		.07	

Note. The full model allows d' to differ between the two age groups being compared. The constrained model holds d' constant across the two age groups being compared. Overall χ^2, df and p rows represent goodness-of-fit statistics when the model was fit to the two age groups together.

Next, we compared performance on the fair and unfair lineups within each age group. The observed and predicted data for the fair lineups are shown in Table 3.2 in Chapter 3, whereas the observed and predicted data for the unfair lineups are shown in Table D.1. We performed a separate fair versus unfair pairwise comparison for the young, middle-aged and older adults. Again, we first fit the model allowing d' to differ across the fair and unfair conditions and then fit the same model constraining d' to be equal across the fair and unfair lineups. The overall χ^2, df and p rows in Table D.3 show the full (unconstrained) and constrained model fit statistics. In comparison to the full model, the constrained model provided a significantly worse fit of the data for the young adults, χ^2 (1) = 5.13, p = .02, and middle-aged adults, χ^2 (1) = 8.67, p = .003, but this did not reach statistical significance in the older adults, χ^2 (1) = 3.03, p = .08. These results indicate that there was a statistically significant difference in d' between the fair and unfair lineups in the young and middle-aged adults, but this only approached significance in the older adults (p = .08). Although the difference in discriminability between the fair and unfair lineups in the older adults was marginally significant in the ROC analysis (p = .05), descriptively speaking, the modelling results were consistent with our ROC results. Both analyses suggest that all three age groups were less able to distinguish between innocent and guilty suspects in the unfair lineups than in the fair lineups.

Table D.3
Full and Constrained (d') Model Fits for the Fair vs. Unfair Lineup Comparisons in the Young, Middle-aged, and Older Adults

	Young				Middle-aged				Older			
	Full model		Constrained model		Full model		Constrained model		Full model		Constrained model	
Estimate	Fair	Unfair	Fair	Unfair	Fair	Unfair	Fair	Unfair	Fair	Unfair	Fair	Unfair
μ_{guilty} (d')	1.21	0.83	1.14	1.14	1.07	0.59	0.98	0.98	0.72	0.43	0.66	0.66
μ_{foil}	-	–1.67	-	–1.50	-	–1.44	-	–1.23	-	–1.29	-	–1.16
c_1	1.13	–0.03	1.12	0.14	1.12	–0.12	1.11	0.08	1.04	–0.05	1.03	0.08
c_2	1.31	0.12	1.29	0.28	1.28	0.00	1.26	0.20	1.21	0.19	1.20	0.31
c_3	1.54	0.28	1.53	0.45	1.44	0.17	1.43	0.36	1.40	0.39	1.39	0.51
c_4	1.89	0.76	1.88	0.94	1.86	0.78	1.84	0.99	1.92	0.88	1.91	1.00
c_5	2.44	1.14	2.42	1.31	2.39	1.26	2.38	1.47	2.45	1.36	2.44	1.47
Overall χ^2	23.72		28.85		35.44		44.11		25.13		28.16	
Overall df	22		23		22		23		22		23	
Overall p	.36		.19		.03		.01		.29		.21	

Note. The full model allows d' to differ between the two lineups being compared. The constrained model holds d' constant across the two lineups being compared. Overall χ^2, df and p rows represent goodness-of-fit statistics when the model was fit to the two lineups together.

Appendix E: Confidence and accuracy in Chapter 3

Table E.1

Frequencies of Identification Responses Made by the Young, Middle-aged, and Older Adults in Each Confidence Bin in the Fair and Unfair Lineups

Lineup type and confidence	Young						Middle-aged						Older					
	Target present			Target absent			Target present			Target absent			Target present			Target absent		
	Guilty suspect	Foil	Inc. reject	Innocent suspect	Foil	Correct reject	Guilty suspect	Foil	Inc. reject	Innocent suspect	Foil	Correct reject	Guilty suspect	Foil	Inc. reject	Innocent suspect	Foil	Correct reject
Fair																		
0–20	11.00	19.00	19.00	6.67	33.33	18.00	15.00	21.00	12.00	4.33	21.67	22.00	11.00	23.00	19.00	5.83	29.17	18.00
30–40	22.00	28.00	16.00	7.00	35.00	27.00	16.00	13.00	12.00	6.33	31.67	16.00	14.00	25.00	21.00	7.00	35.00	23.00
50–60	38.00	25.00	27.00	9.50	47.50	41.00	40.00	35.00	31.00	12.33	61.67	47.00	31.00	61.00	38.00	14.83	74.17	41.00
70–80	40.00	21.00	17.00	8.33	41.67	39.00	45.00	21.00	23.00	8.00	40.00	28.00	26.00	25.00	17.00	6.67	33.33	29.00
90–100	39.00	10.00	8.00	2.50	12.50	19.00	24.00	15.00	17.00	3.50	17.50	28.00	15.00	6.00	8.00	3.00	15.00	13.00
Unfair																		
0–20	0.00	2.00	0.00	3.00	4.00	3.00	1.00	3.00	4.00	1.00	4.00	4.00	2.00	4.00	5.00	3.00	11.00	4.00
30–40	6.00	1.00	7.00	5.00	1.00	5.00	2.00	2.00	0.00	2.00	7.00	0.00	9.00	3.00	7.00	2.00	6.00	4.00
50–60	25.00	1.00	4.00	11.00	4.00	7.00	28.00	9.00	5.00	17.00	4.00	13.00	18.00	10.00	7.00	10.00	11.00	6.00
70–80	19.00	2.00	4.00	7.00	2.00	10.00	19.00	5.00	5.00	11.00	2.00	8.00	19.00	3.00	5.00	9.00	3.00	3.00
90–100	41.00	0.00	1.00	14.00	2.00	11.00	28.00	0.00	2.00	11.00	2.00	3.00	20.00	1.00	0.00	11.00	0.00	6.00

Note. Inc. reject = incorrect rejection; Correct reject = correct rejection.

Appendix F: Confidence and accuracy in young-old and old-old adults in Chapter 3

Older adults made slightly (but not significantly) less accurate suspect identifications at every level of confidence than did young and middle-aged adults. To investigate this further, we separated our older adults into young-old (aged 60–70, $n = 463$) and old-old (aged 71+, $n = 225$) groups. We constructed confidence-accuracy curves in the same way as in Chapter 3. The frequencies of identification responses in each confidence bin are shown in Table F.1. Figure F.1 shows the confidence-accuracy curves for the fair lineups in the young-old and old-old groups. First, it is important to note that the unexpectedly poor accuracy of the old-old adults at the highest level of confidence should be treated with caution because there were only six subjects in this age group who identified a lineup member with this level of confidence. Focusing on the remaining confidence levels (i.e., 0–20, 30–40, 50–60, 70–80), we can see that, as before, the error bars for each age group overlap. This indicates that the differences in suspect identification accuracy between the age groups at each level of confidence are not statistically reliable. Nevertheless, the same numerical trend that we observed in our main confidence-accuracy analysis is apparent: old-old adults are slightly (but not significantly) less accurate at every level of confidence than the young-old adults. This suggests that as memory ability declines with age, older adults do not adjust their criteria to the extent required for them to be just as accurate at each level of confidence as their younger counterparts.

Table F.1
Frequencies of Identification Responses Made by the Young-old and Old-old Adults in Each Confidence Bin in the Fair Lineups

	Target present			Target absent		
Confidence	Guilty suspect	Foil	Incorrect rejection	Innocent suspect	Foil	Correct rejection
			Young-old			
0–20	8.00	16.00	16.00	4.00	20.00	16.00
30–40	10.00	18.00	16.00	4.67	23.33	11.00
50–60	19.00	34.00	19.00	9.50	47.50	30.00
70–80	20.00	14.00	12.00	4.33	21.67	21.00
90–100	15.00	4.00	7.00	2.33	11.67	8.00
			Old-old			
0–20	3.00	7.00	3.00	1.83	9.17	2.00
30–40	4.00	7.00	5.00	2.33	11.67	12.00
50–60	12.00	27.00	19.00	5.33	26.67	11.00
70–80	6.00	11.00	5.00	2.33	11.67	8.00
90–100	0.00	2.00	1.00	0.67	3.33	5.00

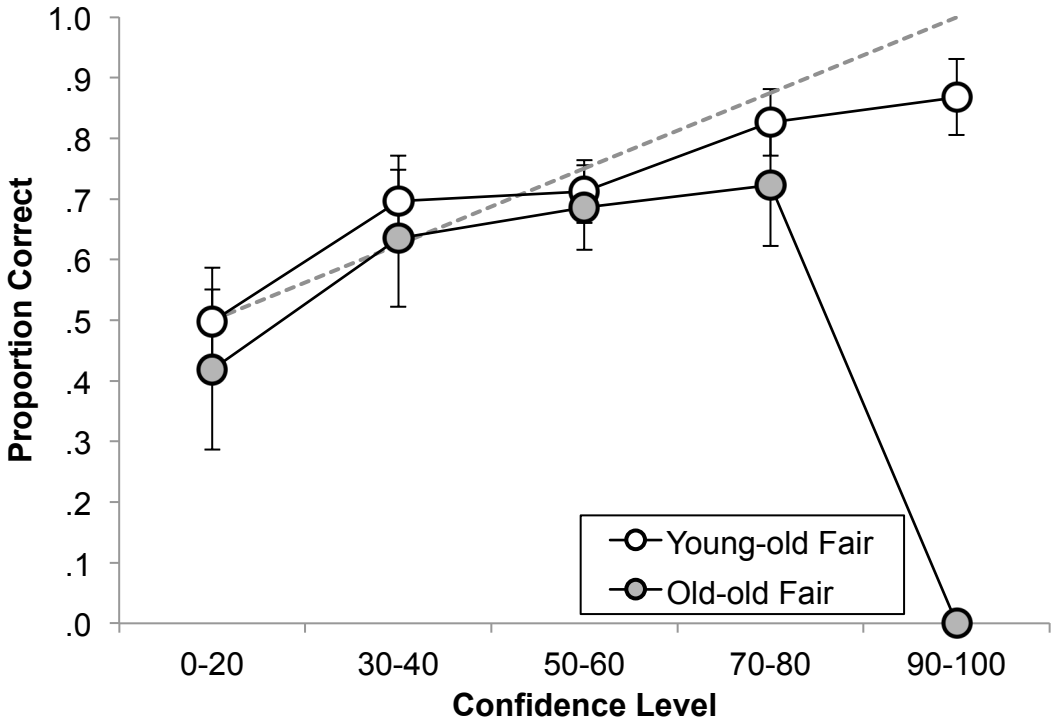

Figure F.1. Confidence-accuracy curves for suspect identifications made by young-old and old-old adults in the fair lineups. Error bars ±1 *SE*. The dashed line represents chance accuracy at the lowest confidence bin (i.e., 0–20) and perfect accuracy at the highest confidence bin (i.e., 90–100).

Appendix G: Modelling in Chapter 4

To confirm our findings from the ROC analysis, we fit a signal detection model to our data (Wixted & Mickes, 2014; see Chapter 1 for a description of the model). We used the same model-fitting procedure outlined in Appendix A, but, this time, we collapsed our data to a 3-point confidence scale: 0–20 (c_1), 30–40 (c_2), 50–60 (c_3). It is more difficult to assume an asymptotic chi-square distribution for cells with fewer than 5 observations (Cochran, 1952). Thus, we chose this 3-point scale because it limited the number of cells with small numbers of observations.[12]

The fair lineup data contained 9 degrees of freedom because there were 3 levels of confidence for the guilty suspect identifications and foil identifications in target-present lineups, and 3 levels of confidence for foil identifications in target-absent lineups. Once these response frequencies were known, the number of lineup rejections was fixed. The model had 4 free parameters (μ_{guilty}, c_1, c_2, c_3) because, for simplicity, we fixed $\mu_{innocent}$ to 0 and set the standard deviations to 1. Thus, the fit for the fair lineups had 9 – 4 = 5 degrees of freedom. The unfair lineup data had 12 degrees of freedom because there were 3 levels of confidence for the guilty suspect identifications and foil identifications in target-present lineups, and 3 levels of confidence for innocent suspect identifications and foil identifications in target-absent lineups. The model had 5 free parameters (μ_{guilty}, μ_{foil}, c_1, c_2, c_3) because, again, we fixed $\mu_{innocent}$ to 0 and set the standard deviations to 1. Thus, the fit for the unfair lineups had 12 – 5 = 7 degrees of freedom.

Table G.1 shows our observed data and the values predicted by the best-fitting model, while Table G.2 shows the best-fitting parameters and the chi-square goodness-of-fit statistics. While it is clear from Table G.1 that this simple model proficiently captured the trends in our data, the (full model) chi-square goodness-of-fit statistics in Table G.2 indicate that some of our data deviated from the predictions of this simple model, suggesting that a more complex model might fit the data better. We first compared performance on the replication, pixelation, block and do-nothing

[12] Some cells in the do-nothing lineups had fewer than 5 observations. Regardless of how the data were collapsed, this was unavoidable because only 4 subjects made foil identifications with greater than 60% confidence after watching the distinctive suspect, and only 5 subjects identified the innocent suspect after watching the non-distinctive culprit.

lineups after subjects had watched the non-distinctive culprit.[13] We performed six separate pairwise comparisons: replication versus pixelation, replication versus block, pixelation versus block, replication versus do-nothing, pixelation versus do-nothing, and block versus do-nothing. We fit the model allowing d' to differ in the two lineups being compared (full model), then we fit the same model, allowing the confidence criteria to differ, but constraining d' to be equal across the two lineups (constrained model). Table G.2 shows the best-fitting parameters and the chi-square goodness-of-fit statistics for the full and constrained models. In comparison to the full model, the constrained model did not provide significantly worse fit of the data for the replication and pixelation, χ^2 (1) = 0.60, p = .44, replication and block, χ^2 (1) = 0.00, p = .96, or pixelation and block, χ^2 (1) = 0.57, p = .45, comparisons. Neither did the constrained model provide a significantly worse fit of the data for the replication and do-nothing, χ^2 (1) = 0.12, p = .73, pixelation and do-nothing, χ^2 (1) = 0.79, p = .37, and block and do-nothing, χ^2 (1) = 0.12, p = .73, comparisons.[14] Consistent with the results of the ROC analysis, this indicates that when the culprit did not have a distinctive feature during crime, all four lineups elicited equivalent discriminability.

Next, we compared performance on the replication, pixelation, block and do-nothing lineups after subjects had watched the distinctive culprit.[15] We used the same procedure: once fitting the model allowing d' to differ in the two lineups being compared (full model), then fitting the same model but constraining d' to be equal across the two lineups (constrained model). Again, Table G.2 shows the best-fitting parameters and the chi-square goodness-of-fit statistics for the full and constrained models. It is clear from Table G.2 (full model) that, compared to the fair (replication, pixelation and block) lineups, doing nothing to prevent the distinctive suspect from

---

[13] When we fit the data to the replication, pixelation, block and do-nothing lineup data separately in subjects who had watched a non-distinctive culprit, the model fit the pixelation (p = .10) and do-nothing (p = .14) data well, but significantly deviated from the observed data in the replication (p = .02) and block (p = .005) conditions.

[14] The fit of the model in the block condition—but not the replication, pixelation or do-nothing conditions—was significantly improved by allowing for unequal variance (i.e., allowing σ_{guilty} to vary). When we conducted the same model-fitting analysis but allowed for unequal-variance in the block condition, we found the same results.

[15] When we fit the data to the replication, pixelation, block and do-nothing lineup data separately in subjects who had watched a distinctive culprit, the model fit the replication (p = .66) and block (p = .11) data well, but significantly deviated from the observed data in the pixelation (p < .001) and do-nothing (p = .01) conditions.

standing out impaired *d′*. Indeed, the constrained model did not provide a significantly worse fit of the data for the replication and pixelation, χ^2 (1) = 1.28, p = .26, replication and block, χ^2 (1) = 0.51, p = .47, or pixelation and block, χ^2 (1) = 0.20, p = .65, comparisons. But, the constrained model did provide a significantly worse fit of the data for the replication and do-nothing, χ^2 (1) = 28.03, p < .001, pixelation and do-nothing, χ^2 (1) = 18.05, p < .001, and block and do-nothing, χ^2 (1) = 22.19, p < .001, comparisons.[16] Again, consistent with the results of the ROC analysis, this indicates that when the culprit had a distinctive feature during crime, all three fair lineups were equally effective, and all three fair techniques enhanced subjects' ability to discriminate between innocent and guilty suspects more than doing nothing to prevent the distinctive suspect from standing out.

Finally, we compared the identification performance of subjects who had watched the non-distinctive culprit with subjects who had watched the distinctive culprit on each lineup type. We performed four separate pairwise comparisons, one for each lineup type, using the same model-fitting procedure. Table G.3 shows the best-fitting parameters and the chi-square goodness-of-fit statistics for the full and constrained models. It is clear from Table G.3 (full model) that doing nothing to prevent the distinctive suspect from standing out impaired *d′* when subjects had watched a distinctive, but not a non-distinctive, culprit. Indeed, in comparison to the full model, the constrained model did not provide significantly worse fit of the distinctive and non-distinctive culprit data for the replication, χ^2 (1) = 0.96, p = .33, pixelation, χ^2 (1) = 0.37, p = .54, and block, χ^2 (1) = 0.09, p = .76, lineups. However, the constrained model did provide a significantly worse fit of the distinctive and non-distinctive culprit data for the do-nothing lineup, χ^2 (2) = 105.80, p < .001. Again, replicating the results of our ROC analysis, this suggests that subjects' ability to distinguish between innocent and guilty suspects on the fair lineups was similar, regardless of whether the culprit had the feature during the crime, but ability to discriminate between innocent and guilty suspects on unfair lineups was better when subjects did not have the opportunity to encode the feature during the crime, compared to when subjects had a memory of that feature.

---

[16] The fit of the model in the pixelation condition—but not the replication, block or do-nothing conditions—was significantly improved by allowing for unequal variance (i.e., allowing σ_{guilty} to vary). When we conducted the same model-fitting analysis but allowed for unequal-variance in the pixelation condition, we found the same results.

Table G.1

Observed and Predicted Identification Responses in Each Confidence Bin in the Replication, Pixelation, Block, and Do-nothing lineups for the Non-distinctive and Distinctive Culprits

	Replication					Pixelation					Block					Do-nothing					
	Target present			Target absent		Target present			Target absent		Target present			Target absent		Target present			Target absent		
Confidence	Guilty suspect	Foil	Inc. reject	Foil	Correct reject	Guilty suspect	Foil	Inc. reject	Foil	Correct reject	Guilty suspect	Foil	Inc. reject	Foil	Correct reject	Guilty suspect	Foil	Inc. reject	Innocent suspect	Foil	Correct reject
Non-distinctive culprit																					
0–50																					
Observed	34.00	16.00	-	42.00	-	33.00	30.00	-	45.00	-	26.00	24.00	-	44.00	-	24.00	29.00	-	1.00	44.00	-
Predicted	22.85	21.13	-	48.22	-	25.36	26.39	-	57.09	-	22.92	21.44	-	49.17	-	24.61	24.99	-	6.27	43.57	-
60–70																					
Observed	24.00	11.00	-	37.00	-	22.00	12.00	-	31.00	-	40.00	8.00	-	28.00	-	25.00	10.00	-	1.00	21.00	-
Predicted	25.74	16.01	-	28.33	-	21.74	15.39	-	25.98	-	28.14	17.32	-	30.44	-	20.13	14.36	-	2.68	19.96	-
80–100																					
Observed	44.00	11.00	-	21.00	-	39.00	11.00	-	24.00	-	36.00	10.00	-	24.00	-	39.00	13.00	-	4.00	18.00	-
Predicted	45.51	12.37	-	17.63	-	41.14	13.80	-	19.30	-	42.90	11.27	-	15.90	-	40.94	14.28		2.00	16.59	-
Total																					
Observed	-	-	44.00	-	84.00	-	-	33.00	-	80.00	-	-	39.00	-	87.00	-	-	46.00	-	-	97.00
Predicted	-	-	40.39	-	89.82	-	-	36.19	-	77.63	-	-	39.00	-	87.49	-	-	46.68	-	-	94.93
Distinctive culprit																					
0–50																					
Observed	27.00	18.00	-	47.00	-	31.00	20.00	-	43.00	-	32.00	23.00	-	42.00	-	20.00	14.00	-	17.00	23.00	-
Predicted	23.95	19.47	-	48.90	-	22.46	21.70	-	49.70	-	25.07	21.75	-	49.80	-	21.49	9.06	-	26.44	20.18	-
60–70																					
Observed	30.00	12.00	-	23.00	-	32.00	10.00	-	22.00	-	30.00	15.00	-	34.00	-	36.00	4.00	-	19.00	11.00	-
Predicted	25.12	13.36	-	25.18	-	21.49	14.36	-	25.67	-	30.34	16.79	-	29.29	-	28.49	6.75	-	24.37	11.91	-
80–100																					
Observed	48.00	11.00	-	21.00	-	39.00	9.00	-	33.00	-	42.00	4.00	-	18.00	-	90.00	0.00	-	44.00	8.00	-
Predicted	51.90	11.33	-	16.76	-	47.24	14.38	-	20.78	-	42.47	9.70	-	13.59	-	89.24	5.62	-	36.90	7.75	-
Total																					
Observed	-	-	35.00	-	90.00	-	-	37.00	-	80.00	-	-	43.00	-	95.00	-	-	18.00	-	-	60.00
Predicted	-	-	35.87	-	90.16	-	-	36.37	-	81.84	-	-	42.88	-	96.33	-	-	21.34	-	-	54.46

Note. The total row displays all reject identification decisions because the model does not account for the confidence level with which lineup rejections are made. Inc. reject = incorrect rejection; Correct reject = correct rejection.

Table G.2

Full and Constrained (d') Model Fits for the Replication vs. Pixelation, Replication vs. Block, Pixelation vs. Block, Replication vs. Do-nothing, Pixelation vs. Do-nothing, and Block vs. Do-nothing Comparisons for the Non-distinctive and Distinctive Culprits

	Non-distinctive culprit				Distinctive culprit			
	Full model		Constrained model		Full model		Constrained model	
Estimate	Replication	Pixelation	Replication	Pixelation	Replication	Pixelation	Replication	Pixelation
μ_{guilty} (d')	1.47	1.36	1.60	1.45	1.60	1.45	1.52	1.52
c_1	1.21	1.12	1.23	1.17	1.23	1.17	1.21	1.18
c_2	1.68	1.67	1.72	1.65	1.72	1.65	1.70	1.67
c_3	2.13	2.08	2.14	2.04	2.14	2.04	2.12	2.06
Overall χ^2	22.63		23.23		25.21		26.49	
Overall df	10		11		10		11	
Overall p	.01		.02		.005		.005	
	Replication	Block	Replication	Block	Replication	Block	Replication	Block
μ_{guilty} (d')	1.47	1.47	1.60	1.51	1.60	1.51	1.55	1.55
c_1	1.21	1.20	1.23	1.25	1.23	1.25	1.22	1.26
c_2	1.68	1.67	1.72	1.73	1.72	1.73	1.71	1.74
c_3	2.13	2.17	2.14	2.25	2.14	2.25	2.13	2.26
Overall χ^2	30.23		30.23		12.21		12.72	
Overall df	10		11		10		11	
Overall p	.001		.002		.27		.31	
	Pixelation	Block	Pixelation	Block	Pixelation	Block	Pixelation	Block
μ_{guilty} (d')	1.36	1.47	1.45	1.51	1.45	1.51	1.48	1.48
c_1	1.12	1.20	1.17	1.25	1.17	1.25	1.24	1.17
c_2	1.67	1.67	1.65	1.73	1.65	1.73	1.72	1.66
c_3	2.08	2.17	2.04	2.25	2.04	2.25	2.24	2.05
Overall χ^2	26.10		26.67		30.93		31.13	
Overall df	10		11		10		11	
Overall p	.004		.005		.001		.001	

	Replication	Do-nothing	Replication	Do-nothing	Replication	Do-nothing	Replication	Do-nothing
μ_{guilty} (d')	1.47	1.53	1.60	0.81	1.60	0.81	1.27	1.00
μ_{foil}	-	0.20	-	−1.51	-	−1.51	-	−1.29
c_1	1.21	1.42	1.23	−0.11	1.23	−0.11	1.17	0.10
c_2	1.68	1.91	1.72	0.38	1.72	0.38	1.64	0.60
c_3	2.13	2.28	2.14	0.82	2.14	0.82	2.05	1.04
Overall χ^2	24.43		24.55		22.25		50.28	
Overall df	12		13		12		13	
Overall p	.02		.03		.04		<.001	
	Pixelation	Do-nothing	Pixelation	Do-nothing	Pixelation	Do-nothing	Pixelation	Do-nothing
μ_{guilty} (d')	1.36	1.53	1.45	0.81	1.45	0.81	1.18	1.18
μ_{foil}	-	0.20	-	−1.51	-	−1.51	-	−1.33
c_1	1.12	1.42	1.17	−0.11	1.17	−0.11	1.12	0.06
c_2	1.67	1.91	1.65	0.38	1.65	0.38	1.60	0.56
c_3	2.08	2.28	2.04	0.82	2.04	0.82	1.99	1.00
Overall χ^2	20.31		21.10		40.97		59.02	
Overall df	12		13		12		13	
Overall p	.06		.07		<.001		<.001	
	Block	Do-nothing	Block	Do-nothing	Block	Do-nothing	Block	Do-nothing
μ_{guilty} (d')	1.47	1.53	1.51	0.81	1.51	0.81	1.23	1.23
μ_{foil}	-	0.20	-	−1.51	-	−1.51	-	−1.31
c_1	1.20	1.42	1.25	−0.11	1.25	−0.11	1.19	0.08
c_2	1.67	1.91	1.73	0.38	1.73	0.38	1.66	0.58
c_3	2.17	2.28	2.25	0.82	2.25	0.82	2.16	1.02
Overall χ^2	27.91		28.03		27.97		50.16	
Overall df	12		13		12		13	
Overall p	.006		.009		.006		<.001	

Note. The full (unconstrained) model allows d' to differ between the two lineups being compared. The constrained model holds d' constant across the two lineups being compared. Overall χ^2, df and p rows represent goodness-of-fit statistics when the model was fit to the two lineups together.

Table G.3

Full and Constrained (d') Model fits for the Distinctive vs. Non-distinctive Comparisons in the Replication, Pixelation, Block, and Do-nothing Lineups

	Full model		Constrained model	
Estimate	Non-distinctive culprit	Distinctive culprit	Non-distinctive culprit	Distinctive culprit
		Replication		
μ_{guilty} (d')	1.47	1.60	1.53	1.53
c_1	1.21	1.23	1.23	1.22
c_2	1.68	1.72	1.69	1.70
c_3	2.13	2.14	2.14	2.13
Overall χ^2	16.62		17.58	
Overall df	10		11	
Overall p	.08		.09	
		Pixelation		
μ_{guilty} (d')	1.36	1.45	1.40	1.40
c_1	1.12	1.17	1.13	1.16
c_2	1.67	1.65	1.68	1.64
c_3	2.08	2.04	2.09	2.04
Overall χ^2	31.22		31.59	
Overall df	10		11	
Overall p	.001		.001	
		Block		
μ_{guilty} (d')	1.47	1.51	1.49	1.49
c_1	1.20	1.25	1.20	1.24
c_2	1.67	1.73	1.67	1.72
c_3	2.17	2.25	2.17	2.24
Overall χ^2	25.82		25.91	
Overall df	10		11	
Overall p	.004		.007	
		Do-nothing		
μ_{guilty} (d')	1.53	0.81	0.85	0.85
μ_{foil}	0.20	–1.51	–0.87	–0.87
c_1	1.42	–0.11	0.48	0.25
c_2	1.91	0.38	1.00	0.65
c_3	2.28	0.82	1.37	1.06
Overall χ^2	30.06		135.86	
Overall df	14		16	
Overall p	.008		<.001	

Note. The full model allows d' to differ between the two conditions being compared. The constrained model holds d' constant across the two conditions being compared. Overall χ^2, df and p rows represent goodness-of-fit statistics when the model was fit to the two conditions together.

Appendix H: Confidence and accuracy in Chapter 4

Table H.1

Frequencies of Identification Responses Made in Each Confidence Bin in the Replication, Pixelation, Block, and Do-nothing Lineups for the Non-distinctive and Distinctive Culprits

	Non-distinctive culprit						Distinctive culprit					
Lineup type	Target present			Target absent			Target present			Target absent		
and confidence	Guilty suspect	Foil	Incorrect rejection	Innocent suspect	Foil	Correct rejection	Guilty suspect	Foil	Incorrect rejection	Innocent suspect	Foil	Correct rejection
Replication												
0–20	8.00	2.00	2.00	1.67	8.33	7.00	5.00	4.00	3.00	1.33	6.67	7.00
30–40	11.00	10.00	7.00	2.50	12.50	9.00	9.00	9.00	9.00	3.50	17.50	11.00
50–60	23.00	12.00	11.00	6.33	31.67	18.00	33.00	10.00	9.00	5.33	26.67	27.00
70–80	32.00	10.00	10.00	3.83	19.17	30.00	34.00	14.00	8.00	3.67	18.33	28.00
90–100	28.00	4.00	14.00	2.33	11.67	20.00	24.00	4.00	6.00	1.33	6.67	17.00
Pixelation												
0–20	5.00	5.00	0.00	1.33	6.67	6.00	3.00	3.00	3.00	1.17	5.83	5.00
30–40	12.00	11.00	6.00	3.83	19.17	11.00	12.00	7.00	2.00	3.17	15.83	14.00
50–60	27.00	21.00	15.00	4.83	24.17	18.00	32.00	18.00	13.00	5.17	25.83	16.00
70–80	32.00	11.00	8.00	4.50	22.50	29.00	29.00	7.00	12.00	3.50	17.50	24.00
90–100	18.00	5.00	4.00	2.17	10.83	16.00	26.00	4.00	7.00	3.33	16.67	21.00
Block												
0–20	6.00	6.00	4.00	1.67	8.33	7.00	4.00	9.00	3.00	2.17	10.83	8.00
30–40	9.00	10.00	8.00	2.50	12.50	11.00	12.00	8.00	8.00	2.83	14.17	12.00
50–60	28.00	15.00	9.00	6.50	32.50	31.00	28.00	12.00	12.00	5.17	25.83	27.00
70–80	39.00	6.00	14.00	3.17	15.83	19.00	29.00	10.00	11.00	4.17	20.83	34.00
90–100	20.00	5.00	4.00	2.17	10.83	19.00	31.00	3.00	9.00	1.33	6.67	14.00
Do-nothing												
0–20	4.00	7.00	4.00	0.00	14.00	6.00	2.00	2.00	0.00	1.17	5.83	8.00
30–40	8.00	12.00	7.00	0.00	14.00	13.00	5.00	9.00	3.00	1.33	6.67	3.00
50–60	20.00	14.00	15.00	2.00	23.00	27.00	34.00	6.00	6.00	2.00	10.00	18.00
70–80	33.00	14.00	12.00	4.00	24.00	26.00	36.00	1.00	7.00	2.00	10.00	14.00
90–100	23.00	5.00	8.00	0.00	8.00	25.00	69.00	0.00	2.00	0.50	2.50	17.00

Appendix I: Confidence and accuracy in Chapter 5

Table I.1
Frequencies of Identification Responses in Each Confidence Bin in the Low-Variation, Moderate-Variation, and Do-nothing Lineups in Experiment 1

	Target present			Target absent		
Lineup type and confidence	Guilty suspect	Foil	Incorrect rejection	Innocent suspect	Foil	Correct rejection
Low-variation						
0–60	111.00	75.00	73.00	57.00	110.00	88.00
70–80	58.00	24.00	37.00	25.00	48.00	57.00
90–100	51.00	12.00	25.00	12.00	22.00	47.00
Moderate-variation						
0–60	107.00	53.00	71.00	56.00	78.00	110.00
70–80	81.00	17.00	30.00	42.00	24.00	65.00
90–100	62.00	6.00	26.00	14.00	14.00	50.00
Do-nothing						
0–60	103.00	48.00	43.00	71.00	68.00	74.00
70–80	92.00	13.00	24.00	65.00	20.00	43.00
90–100	117.00	4.00	20.00	81.00	10.00	32.00

Table I.2
Frequencies of Identification Responses in Each Confidence Bin in the Low-Variation, High-Variation, and Do-nothing Lineups in Experiment 2

	Target present			Target absent		
Lineup type and confidence	Guilty suspect	Foil	Incorrect rejection	Innocent suspect	Foil	Correct rejection
Low-variation						
0–60	112.00	75.00	73.00	29.00	129.00	121.00
70–80	62.00	22.00	32.00	11.00	36.00	65.00
90–100	62.00	11.00	18.00	5.00	20.00	51.00
High-variation						
0–60	106.00	23.00	48.00	68.00	48.00	98.00
70–80	118.00	6.00	30.00	24.00	8.00	91.00
90–100	119.00	5.00	17.00	29.00	9.00	97.00
Do-nothing						
0–60	125.00	37.00	54.00	72.00	70.00	105.00
70–80	80.00	9.00	20.00	37.00	26.00	63.00
90–100	134.00	1.00	9.00	40.00	8.00	48.00